Assessment of MERRA-2 Land Surface Energy Flux Estimates

Clara S. Draper *

USRA/GESTAR and NASA Global Modeling and Assimilation Office, Greenbelt, MD, USA.

Now at CIRES NOAA/ESRL, Physical Sciences Division, Boulder, CO.

Rolf H. Reichle

NASA Global Modeling and Assimilation Office,Greenbelt, MD, USA.

Randal D. Koster

NASA Global Modeling and Assimilation Office,Greenbelt, MD, USA.

* *Corresponding author address:* Clara Draper, NOAA ESRL, Physical Sciences Division, 325 Broadway, Boulder, CO, USA.

E-mail: clara.draper@noaa.gov

ABSTRACT

In the Modern-Era Retrospective analysis for Research and Applications, version 2 (MERRA-2) system the land is forced by replacing the model-generated precipitation with observed precipitation before it reaches the surface. This approach is motivated by the expectation that the resultant improvements in soil moisture will lead to improved land surface latent heating (LH). Here we assess aspects of the MERRA-2 land surface energy budget and 2 m air temperatures (T^{2m}). For global land annual averages, MERRA-2 appears to overestimate the LH (by 5 Wm^{-2}), the sensible heating (by 6 Wm^{-2}), and the downwelling shortwave radiation (by 14 Wm^{-2}), while underestimating the downwelling and upwelling (absolute) longwave radiation (by 10-15 Wm^{-2} each). These results differ only slightly from those for NASA's previous reanalysis, MERRA. Comparison to various gridded reference data sets over Boreal summer (June-July-August) suggests that MERRA-2 has particularly large positive biases (>20 Wm^{-2}) where LH is energy-limited, and that these biases are associated with evaporative fraction biases rather than radiation biases. For time series of monthly means during Boreal summer, the globally averaged anomaly correlations (R_{anom}) with reference data were improved from MERRA to MERRA-2, for LH (from 0.39 to 0.48 vs. GLEAM data) and the daily maximum T^{2m} (from 0.69 to 0.75 vs. CRU data). In regions where T^{2m} is particularly sensitive to the precipitation corrections (including the central US, the Sahel, and parts of south Asia), the changes in the T^{2m} R_{anom} are relatively large, suggesting that the observed precipitation influenced the T^{2m} performance.

1. Introduction

The NASA Global Modeling and Assimilation Office recently released the Modern-Era Retrospective analysis for Research and Applications, version 2 (MERRA-2; Gelaro and Coauthors (2017)). This new global reanalysis product replaces and extends the original MERRA atmospheric reanalysis (Rienecker et al. 2011), as well as the MERRA-Land reanalysis (Reichle et al. 2011). In addition to several other major advances, MERRA-2 uses observed precipitation in place of model-generated precipitation at the land surface during the atmospheric model integration. The use of observed precipitation in MERRA-2 was refined from the approach used for MERRA-Land (Reichle et al. 2017b), which was an offline (land only) replay of MERRA forced by atmospheric fields from MERRA but with the precipitation forcing corrected using gauge-based observations.

The motivation for using observed precipitation in reanalyses is that precipitation is the main driver of soil moisture, which in turn controls the partitioning of incident surface radiation between latent heat (LH) and sensible heat (SH) fluxes back to the atmosphere. Reichle et al. (2017a) show that both MERRA-2 and MERRA-Land have improved upon the land surface hydrology of MERRA, showing better agreement with independent observational time series of soil moisture, terrestrial water storage, stream flow, and snow amount. Here, we extend this work, by evaluating the MERRA-2 surface energy budget and 2 m temperatures (T^{2m}) over land. In particular, we focus on whether the improved hydrology in both the (offline) MERRA-Land and the (coupled land/atmosphere) MERRA-2 data sets translates into the expected improvements to the monthly mean LH and SH. We also expand previous work by evaluating the reanalyses land surface output globally, rather than focusing on locations with high quality ground-based observations.

We start by comparing the long-term annual global energy budget over land from MERRA-2, MERRA-Land, and MERRA to state of the art estimates from the literature. These literature estimates, from Trenberth et al. (2009), Wild et al. (2015), and the NASA Energy and Water Cycle Studies program (NEWS,NSIT (2007); L'Ecuyer et al. (2015)) were each produced by carefully combining multiple input data sets with global energy balance constraints. Taken together they represent our best understanding of the long-term annual mean energy budget over land.

Next, we consider global maps of the performance of the land surface turbulent heat fluxes from each reanalyses, as a step towards linking differences in performance to the dominant local physical processes and to the potential improvements obtained from the use of the observed precipitation in MERRA-2. We focus on the Boreal summer (June-July-August; JJA), since land/atmosphere coupling is strongest and surface turbulent heat fluxes are most active in the summer.

Unfortunately, there are no standard global gridded reference data sets against which the reanalysis LH and SH can be evaluated. Several recent efforts have compared global LH estimates from different combinations of reanalyses, offline land surface models, and diagnostic methods. Most estimates generally agree on the regional patterns and local seasonal cycle of LH, although there is considerable disagreement in the absolute values and temporal behavior across different flux estimates (Jiménez et al. 2011; Mueller et al. 2011; Miralles et al. 2011). Additionally, uncertainty in the basic model structure is the largest source of disagreement (Schlosser and Gao 2010; Mueller et al. 2013). While ground-based observations are available from tower-mounted eddy covariance sensors (e.g., Baldocchi and Coauthors (2001)), the number of towers (in the 100's) is well below the sampling needed for global estimation (and their locations are not designed to sample globally-representative land cover types). Additionally, the measurements themselves have considerable uncertainty and limited spatial representativeness (up to 1 km).

In the absence of a standard reference, we compare the JJA reanalysis turbulent heat flux estimates to two different gridded reference data sets: Global Land surface Evaporation: the Amsterdam Methodology (GLEAM) (Miralles et al. 2011; Martens et al. 2017) for LH, and Fluxnet-

Model Tree Ensembles (MTE) (Jung et al. 2010) for LH and SH. These data sets were selected for several reasons: i) they are amongst the state of the art, ii) they are available globally for multi-decadal time periods, iii) they are independent of each other, and iv) they rely on very different estimation methodologies (water balance modeling for GLEAM, and upscaling of tower measurements for MTE). Since neither GLEAM nor MTE represents direct observations of the turbulent heat fluxes, we also compare each reanalysis to tower-based eddy covariance observations from the Fluxnet-2015 data set (Fluxnet 2015). To determine the potential contribution of radiation biases to regional LH and SH biases, we also compare the reanalyses surface radiation fields for JJA against gridded observations from the Clouds and the Earth's Radiant Energy System (CERES) and Energy Balanced and Filled (EBAF) data set (Kato et al. 2013).

Finally, to test whether the changes in the surface energy budget from MERRA to MERRA-2 have affected the atmospheric boundary layer, we also evaluate the JJA monthly mean daily minimum and maximum T^{2m} against observations from the Climatic Research Unit (CRU) at the University of East Anglia (Harris et al. 2014). Improvements in MERRA-2 due to the use of observed precipitation cannot be isolated from the many other advances distinguishing MERRA-2 from MERRA. Consequently, we establish whether the improvements in the surface turbulent fluxes and T^{2m} are at least consistent with the expected improvements from the use of observed precipitation, by cross-referencing the evaluation results against the regional sensitivity to precipitation and/or soil moisture.

This paper is organized as follows. Section 2 summarizes the reanalysis and reference data sets used, and Section 3 presents the results, including evaluation of the i) reanalyses annual global land energy budget averages, ii) the spatially distributed mean JJA energy budget and T^{2m}, and ii) the temporal behavior of the JJA turbulent heat fluxes and T^{2m}. We also identify regions of sensitivity to the observed precipitation forcing in MERRA-2, for cross-reference against the evaluation results. Our findings are summarized in Section 4.

2. Methodology and data

a. The reanalyses

The coverage and resolution of each reanalysis is summarized in Table 1, with further details below. MERRA (Rienecker et al. 2011) and MERRA-2 (Gelaro and Coauthors 2017) are atmospheric reanalyses produced with the NASA Goddard Earth Observing System Version 5 (GEOS-5) modeling and data assimilation system, and were designed to provide historical analyses of the hydrological cycle across a broad range of climate time scales. To address shortcomings in the MERRA land surface hydrology, MERRA-Land (Reichle et al. 2011) was released as an offline (land only) replay of MERRA, with the model-generated precipitation corrected using rain-gauge observations and with minor, but important, model parameter changes. MERRA-2 features several major advances from MERRA, including an updated atmospheric general circulation model, an updated atmospheric assimilation system, an interactive aerosol scheme, and the use of observed precipitation at the land surface (and to compute wet aerosol deposition). In addition to the land model updates from MERRA-Land, MERRA-2 includes several more updates relevant to the land, as outlined in Reichle et al. (2017a). Most notably, the surface turbulence scheme was revised, generally resulting in enhanced SH over land (Molod et al. 2015).

The method used to apply the observed precipitation at the land surface in MERRA-2 was refined from that used in MERRA-Land (Reichle and Liu 2014; Reichle et al. 2017b). In MERRA-Land the precipitation was corrected with daily Climate Prediction Center (CPC) Unified (CPCU; Chen et al. (2008)) precipitation observations everywhere. For MERRA-2 the input precipitation differs in two ways: i) in the high latitudes the MERRA-2 model-generated precipitation is retained, and ii) over Africa the MERRA-2 precipitation is corrected with pentad-scale blended satellite and gauge-based observations from the CPC Merged Analysis of Precipitation (CMAP; Xie and Arkin (1997)) and the Global Precipitation Climatology Project (GPCP; Huffman et al. (2009)) version 2.1.

The land surface turbulent fluxes from the NASA reanalyses (MERRA-2, MERRA-Land, and MERRA) have not been explicitly evaluated globally. However, Jiménez et al. (2011) and Mueller et al. (2011) both included MERRA LH when merging multiple LH global land data sets into a single enhanced estimate (see Section 2.b), and in both studies MERRA was amongst the highest of the input LH estimates used. Additionally, Jiménez et al. (2011) noted a sharp gradient in the MERRA LH around 10°S in the tropics that was not present in other LH estimates. This bias gradient was traced to MERRA's excessive rainfall canopy interception and precipitation errors (Reichle et al. 2011). Consequently, the interception reservoir parameters were revised for MERRA-Land (and MERRA-2) to eliminate this feature (the interception reservoir update was the most significant modeling change from MERRA to MERRA-Land).

An additional reanalysis, ERA-Interim, from the European Centre for Medium Range Weather Forecasting (Dee et al. 2011), is included in the evaluation of the temporal behavior of the turbulent fluxes. In contrast to the NASA reanalyses, ERA-Interim includes a land surface updating scheme (de Rosnay et al. 2014). Specifically, the soil moisture, soil temperature, and snow temperatures are updated to minimize errors in the forecast screen-level relative humidity and temperature, while the snow depths are updated using satellite- and ground-based snow cover and snow depth observations.

b. Annual global land energy budget estimates

We compare the reanalyses annual global land energy budgets to three state of the art estimates, from Trenberth et al. (2009), Wild et al. (2015), and the NEWS program estimates of L'Ecuyer et al. (2015). Each of these is based on a weighted merger of multiple modeled and observed data sets, and each applies to the energy budget at the start of the 21st Century. For Trenberth et al. (2009) we have used their estimates for the 'CERES period' of 2000-2004; Wild et al. (2015) nominally refers to the same period; while L'Ecuyer et al. (2015) nominally refers to 2000-2009. Note that the MERRA LH and SH over land were used as one of the inputs in NEWS.

These three global energy budget studies all provide continental and oceanic energy estimates, where 'continental' is defined as non-ocean, and so includes land, land-ice, and lakes, but excludes inland seas. By contrast, the land estimates from MERRA-2, MERRA-Land, and MERRA apply to the area modeled by the land surface model, excluding land-ice, lakes, and inland seas. The discrepancy due to the inclusion or exclusion of land-ice is significant: land-ice accounts for 10% of the continental area, with Antarctica making up 95% of this. NEWS provides energy budgets for each continent separately (L'Ecuyer et al. 2015), and we use their (balance-constrained) energy budget estimates to approximate the land-only energy budget terms by subtracting the area-weighted Antarctica estimates from the global continental estimates. We then use our land-only NEWS estimates to approximate the continental to land ratio for each NEWS energy budget term. By assuming that the same ratios apply to Trenberth et al. (2009) and Wild et al. (2015) we then approximate land-only estimates for the latter two studies. L'Ecuyer et al. (2015) and Wild et al. (2015) both provide uncertainty ranges for their globally averaged continental estimates, which we have applied unchanged to our approximated land-only estimates.

For LH, we have also used three additional global land annual average estimates from the hydrology community, from Jiménez et al. (2011), Mueller et al. (2011), and Mueller et al. (2013). These estimates are also based on merging modeled and observed estimates. Jiménez et al. (2011) applies to global land (using a similar land definition to the NASA reanalyses) for 1994, while Mueller et al. (2011) applies to the global land area, excluding the Sahara, from 1989-1995, and Mueller et al. (2013) applies to the global land plus Greenland for 1989-2005. As previously noted, MERRA LH was one of the inputs used in the multi-product mergers of Jiménez et al. (2011) and Mueller et al. (2011).

c. Gridded reference data sets

The coverage and resolution of each gridded reference data set, together with a brief summary of important interdependencies with other data sets or reanalyses used in the study and uncertainty estimates (where available) are summarized in Table 2, with further details provided below.

1) GLEAM

GLEAM (version 3.1a) provides daily estimates of terrestrial evapotranspiration, estimated from satellite and reanalysis forcing using a Priestley and Taylor-based model (Miralles et al. 2011; Martens et al. 2017). The precipitation is from the Multi-Source Weighted-Ensemble Precipitation, which is a multi-model merger of established precipitation data sets, including the same CPCU data set used in MERRA-Land and MERRA-2, as well as ERA-Interim precipitation (the latter is used predominantly in the high latitudes, where observed precipitation data sets are more uncertain (Beck et al. 2017)). The net surface radiation and T^{2m} are from ERA-Interim. Compared to independent observations from 91 flux towers, GLEAM has an average unbiased root mean square error (ubRMSE; or error standard deviation) of 20 Wm^{-2} and an average anomaly correlation of 0.42 (Martens et al. 2017).

2) MTE

MTE provides global estimates of carbon dioxide, energy, and water fluxes at the land surface, calculated using a machine learning technique to upscale half-hourly energy balance-corrected eddy covariance observations from 253 Fluxnet tower observations (Jung et al. 2011). The input Fluxnet observations are from the La Thuile data release, an earlier generation of the Fluxnet-2015 data set used here (to be introduced in Section 2.d). CPCU precipitation (again, used directly in MERRA-Land and MERRA-2) and a T^{2m} data set based on CRU data (Jung et al. 2011) are used as predictive (regression) variables in the MTE. However, this meteorological data has little impact on the MTE monthly anomalies, which are instead driven by the vegetation variability as observed by the fraction of absorbed Photosynthetically Active Radiation (fPAR; Jung et al. (2010)). When 20% of the Fluxnet training data was withheld from the algorithm, the average Root Mean Square Error (RMSE) with the withheld data was 15 Wm^{-2}, for both LH and SH, and the average anomaly correlation was 0.57 for LH and 0.60 for SH (Jung et al. 2011). In general, the MTE method is better suited to estimating spatial variability and the seasonal cycle than it is to capturing interannual anomaly patterns (Jung et al. 2009).

3) CRU TEMPERATURE DATA

CRU TSv4.00 provides gridded monthly means of the daily mean, minimum, and maximum temperature over land (Harris et al. 2014; University of East Anglia Climate Research Unit et al. 2014). The temperatures are calculated from quality controlled climate station data, which are interpolated onto the grid according to an assumed correlation decay distance (set to 1200 km for temperature variables). In instances where no station data are available within the assumed decay distance, the published data value defaults to the climatology. Here, such climatological values have been screened out. Also, we require at least 10 data points to estimate each statistic for a given grid cell. Even with this screening, the gridded output will be much less certain when/where station coverage is less dense, which occurs over Africa, South America, central Australia, and the high latitudes.

4) CERES-EBAF radiation data

CERES-EBAF version 4.00 surface radiances are produced with a radiative transfer model after adjusting modeled and observed input data for consistency with Top of Atmosphere (TOA) CERES-EBAF radiation (Kato et al. 2013). The input data (surface, cloud, and atmospheric properties) are adjusted according to their observation-based estimated uncertainties. The input temperature and humidity profiles and land surface skin temperature (T_{skin}) are from NASA's GEOS-5.4.1 modeling and assimilation system, the same system (although a different version) used in MERRA and MERRA-2.

The CERES output shortwave irradiances are primarily determined by (observation-based) TOA radiation and clouds, hence they are reasonably independent of the MERRA and MERRA-2 reanalyses (Kato et al. 2013). On the other hand, the CERES output longwave irradiances, and particularly the upwelling longwave (LW_u), are strongly dependent on the GEOS-5 T_{skin} input. However, the CERES algorithm does adjust its input GEOS-5 T_{skin} with observation-based cloud information, so comparison between the CERES-EBAF and GEOS-5 LW_u partly reflects these observation-based adjustments, even though the two fields are not independent. Compared to independent ground-based observations from 24 sites over land, the RMSE of the CERES-EBAF radiation is 12 Wm^{-2} for downwelling shortwave (SW_d), and 10 Wm^{-2} for downwelling longwave (LW_d) (CERES-EBAF 2017). For the regional estimates over land, CERES-EBAF (2017) estimated the uncertainty to be 12 Wm^{-2} for SW_d, 4 Wm^{-2} for upwelling shortwave (SW_u), 10 Wm^{-2} for LW_d, and 18 Wm^{-2} for LW_u.

5) Gridded data set processing

As noted in Tables 1 and 2 some of the reference data sets and reanalyses used here publish output that applies only to the land fraction within each grid cell, while others publish a single estimate that applies to all surface types (land, permanent land-ice, lakes, ocean) within each grid cell. All of the gridded data sets and reanalyses were screened by removing all grid cells where the MERRA-2 land fraction was less than 50% (after interpolation to the relevant resolution), and then aggregated up to monthly means and 1° spatial resolution. All maps of global statistics are based on the Boreal summer months of JJA only, and each comparison is made over the maximum available co-incident time period, with the time periods noted in the relevant figure captions. The anomaly correlations (R_{anom}) are evaluated based on anomalies from the mean seasonal cycle (calculated by subtracting the time period mean separately for each calendar month). The gridded reference data sets were also used to estimate the annual global land average values, for which the (interpolated) MERRA-2 land area in each grid cell was used.

d. Fluxnet-2015 tower observations

The Fluxnet-2015 (Fluxnet 2015) sites were selected by downloading all Tier 1 observations at non-irrigated sites within grid cells classified as land at 1° resolution (as derived above in Section 2.c.5), and for which at least a 10 year data record is available. Eddy covariance sensors underestimate turbulent heat fluxes and do not generally close the energy balance (Wilson et al. 2002), hence we used the Fluxnet-2015 energy balance closure-corrected LH and SH (see Fluxnet (2015) for details of the correction method). While these corrections are rather uncertain, the corrected LH and SH showed better agreement with all of the reanalyses in Table 1 in terms of the means across all sites and the correlation of the means between the sites (while having negligible impact on the mean time series anomaly correlations). The balance-corrected Fluxnet data were then screened to retain only days with less than 10% gap-filled data, and only sites with data for at least 2550 days ($\sim$ 70% of 10 years). The monthly means were then calculated for months with at least 15 days of observations after the above screening, and the corresponding reanalysis monthly means were estimated using the same days. The resulting Fluxnet monthly time series were visually inspected, and obviously unrealistic features were removed. Four sites with unrealistic time series were removed. Of the remaining 21 stations, just one was in the Southern Hemisphere. Since our evaluation focuses on the Boreal summertime, this site was excluded. The remaining 20 sites that have been used in this study are listed in supplemental Table 1.

3. Results

a. Annual global land energy budgets

The globally averaged annual land energy budget estimates for MERRA-2, MERRA-Land, and MERRA are illustrated in Figure 1, with numerical values given in Table 3. For each term, the estimates for MERRA-2 and MERRA are similar (within 2-3 Wm^{-2}), while the partitioning of R_{net} into LH and SH differs for MERRA-Land, which is shifted towards greater SH. Compared to MERRA, MERRA-Land has 11 Wm^{-2} more SH, and 8 Wm^{-2} less LH, with the difference in R_{net} due to decreased LW_u (recall that in the offline MERRA-Land SW_{net} and LW_d are taken directly from MERRA).

Figure 1 also includes the energy budget estimates from the literature (see Section 2.b), as well as the annual global land averages for each of the gridded reference data sets in Table 2. In Figure 1a, the MERRA-2 and MERRA global land LH are higher than all of the other estimates (although MERRA-2 is within the Jiménez et al. (2011) and Wild et al. (2015) confidence intervals). The three (land-adjusted) LH estimates from the global energy budget studies (Trenberth et al. (2009), Wild et al. (2015), and NEWS) are very similar to each other, and to MTE, GLEAM, Mueller et al. (2011), and MERRA-Land (all are within 1 Wm^{-2}). While the other two LH estimates from the hydrology community (Jiménez et al. (2011) and Mueller et al. (2013)) are higher, they are not as high as MERRA-2 and MERRA. Compared to the average of the three global land energy budget estimates, the MERRA-2 LH is biased high by 6 Wm^{-2} (15%), while MERRA is biased high by 9 Wm^{-2} (21%), and MERRA-Land is much closer, being biased high by just 1 Wm^{-2} (2%).

For the global land SH in Figure 1b, MERRA-2 and MERRA are both higher than Trenberth et al. (2009) and Wild et al. (2015), although lower than NEWS (but within the NEWS confidence interval) and very close (within 1 Wm^{-2}) to MTE. Compared to the average of the three global land energy budget estimates, MERRA-2 is biased high by 5 Wm^{-2} (15%) and MERRA by 4 Wm^{-2} (12%), while MERRA-Land is much higher, with a bias of 15 Wm^{-2} (42%).

The positive biases in both LH and SH from the reanalyses indicate a positive bias in the incident energy at the land surface. Indeed, Figure 1g shows that the reanalyses R_{net} exceed the three global energy budget estimates, although MERRA-2 (the lowest of the reanalyses) is only slightly higher (2 Wm^{-2}) than the CERES-EBAF value. Compared to the average of the three global energy budget estimates, the R_{net} biases are 12 Wm^{-2} (15%) for MERRA-2, 13 Wm^{-2} (17%) for MERRA, and 16 Wm^{-2} (21%) for MERRA-Land. Figures 1c-f show that the positive R_{net} bias in MERRA-2 and MERRA is made up of a large positive bias in SW_d combined with insufficient LW_u, both partly offset by underestimated LW_d. For SW_d (Figure 1c) MERRA-2 and MERRA are higher than all three global land energy budget estimates and CERES-EBAF, with a bias compared to the the three-product average of 14 Wm^{-2} (7%) for MERRA-2 and 16 Wm^{-2} (8%) for MERRA. For SW_u (Figure 1d), MERRA-2 and MERRA are both above NEWS, Trenberth et al. (2009), and CERES-EBAF, but below Wild et al. (2015) (although within the confidence interval). Both are biased high by 3 Wm^{-2} (8%), compared to the three-product average. For LW_d (Figure 1e), MERRA-2 and MERRA are lower than the of the other estimates, with biases of -11 Wm^{-2} (-3%) for MERRA-2 and -10 Wm^{-2} (-3%) for MERRA against the three-product average. For LW_u (Figure 1f) MERRA-2, MERRA-Land, and MERRA are again lower than the other plotted estimates, with biases of -11 Wm^{-2} (-3%) for MERRA-2, -13 Wm^{-2} (-3%) for MERRA-Land, and -10 Wm^{-2} (-3%) for MERRA.

The literature estimates in Figure 1 are presented as long term means, and each represents different temporal and spatial coverage. Likewise, the annual global land averages for the gridded reference data sets in Figure 1 are based on the full available (spatial and temporal) coverage for each. However, the gridded reference data sets and reanalyses can be cross-screened to ensure that they are compared with consistent coverage. With this cross-screening, the MERRA-2 LH bias estimate is 7 Wm^{-2} vs. GLEAM, or 9 Wm^{-2}vs. MTE, while the SH bias is 1 Wm^{-2} vs. MTE, and the radiation biases vs. CERES-EBAF are 10 Wm^{-2} for SW_u, 2 Wm^{-2} for SW_d, -18 Wm^{-2} for LW_d, -11Wm^{-2} for LW_u, and $<$0.5 Wm^{-2} for R_{net}. In general, the above-quoted biases (calculated after cross-screening) are all close (within 1 Wm^{-2}) to the values estimated from the data plotted in Figure 1 (which does not include cross-screening), with the exception of the LH bias vs. MTE, which is 6 Wm^{-2} without cross-screening (compared to 9 Wm^{-2}). This discrepancy is due to the MTE global mean being lower than it otherwise would be, due to the lack of coverage over the Sahara (which has near-zero annual mean LH).

b. Land-atmosphere coupling and the MERRA-2 precipitation corrections

Here, we identify regions where, in MERRA-2, i) LH is sensitive to precipitation (or soil moisture), and ii) the daily maximum T^{2m} (T^{2m}_{max}) is sensitive to the applied precipitation corrections. These regions can then be used to determine where the change in performance from MERRA to MERRA-2 is most likely associated with the precipitation corrections. Note that for part ii) above, the diurnal temperature range could be expected to have a stronger signal of the daytime turbulent heat fluxes (Betts et al. 2017), however a preliminary comparison (not shown) revealed similar results for DTR and T^{2m}_{max}, and we have presented the results for T^{2m}_{max} since this variables is included in the published MERRA-2 data sets.

1) Soil Moisture and Latent Heating

To first order, LH (or evapotranspiration) from soil and vegetation surfaces can be conceptualized as either a moisture- or energy-limited process. In drier conditions (i.e., for soil moisture below some critical point), LH is moisture-limited in that it is restricted by the amount of soil moisture available for evapotranspiration. Temporal variations in LH will then be correlated with the plant available soil moisture (principally, the soil moisture in the root-zone). In contrast, in more humid conditions LH is energy limited; there is sufficient soil moisture available for evapotranspiration, so LH proceeds at the maximum rate determined by atmospheric water demand, and temporal variations in LH are accordingly correlated with temporal variations in atmospheric demand (net radiation, atmospheric humidity deficit, and wind), rather than soil moisture.

Figure 2 shows the squared correlation between the JJA monthly anomaly MERRA-2 LH and rootzone soil moisture ($R^2_{anom}(LH,SM)$). Lower $R^2_{anom}(LH,SM)$ indicates a tendency towards energy-limited LH, which for the Boreal summer occurs in the high latitudes, central and eastern Europe, the eastern US, south China, and much of the tropics (the Amazon, equatorial Africa, and southeast Asia). On the other hand, higher $R^2_{anom}(LH, SM)$ indicates a tendency towards moisture-limited LH, and occurs across the remainder of the low and mid-latitudes. While we have plotted JJA to focus on the Boreal summer, there are still regions of moisture-limited LH in the southern hemisphere during Austral winter, specifically in arid regions (southern Africa, much of Australia, and the desert and steppe regions of South America).

2) Precipitation feedback on air temperature

Figure 3 shows maps of the squared anomaly correlation (R^2_{anom}) between anomaly timeseries of JJA MERRA-2 monthly T^{2m}_{max} and anomaly timeseries of 2-month (current + previous month) averaged MERRA-2 precipitation. For example, the June T^{2m}_{max} is compared to the (May+June) precipitation, while the July T^{2m}_{max} is compared to the (June+July) precipitation, and so on. The precipitation is lagged like this to allow the precipitation signal to accumulate in the soil, and influence the subsequent T^{2m}_{max}. In Figure 3a the MERRA-2 model-generated precipitation (PRECTOT) is used, while in Figure 3b the MERRA-2 observation-corrected precipitation (PRECTOTCORR) is used. The R^2_{anom} are plotted only for negative R values, since the dominant local relationship between precipitation and daytime temperature is negative (i.e., under moisture-limited conditions, reduced precipitation leads to reduced soil moisture, which limits LH and increases SH and T^{2m}). Figure 3b reflects the modeled relationship in MERRA-2 between precipitation falling on the surface and T^{2m}_{max}.Even with the difference in time periods, the patterns are similar to those found across the contiguous U.S. from observations by Koster et al. (2015).

Figure 3c then shows the difference between $R^2_{anom}(T^{2m}_{max}, PRECTOTCORR)$ and $R^2_{anom}(T^{2m}_{max}, PRECTOT)$. This difference ($\Delta R^2_{anom}$) is the increase in the fraction of variance in T^{2m}_{max} explained by the (observed) precipitation seen by the land (PRECTOTCORR) over that explained by the model-generated precipitation (PRECTOT). It thus provides a measure of the local impact of the observed precipitation on the MERRA-2 T^{2m}_{max}. This measure is sensitive to both the magnitude of the precipitation corrections and the local response of the atmospheric model to those corrections. Note that the lack of sensitivity in the high latitudes was inevitable for this metric, since the model-generated precipitation is used there.

For the Boreal summer, the strongest impact of the observed precipitation, which can explain more than 25% of the T^{2m}_{max} variance, is indicated in the central US, central America, the northern tip of South America, across a broad swath along the Sahel, and parts of south Asia. Note that these regions do not directly correspond to the regions of strongest moisture-limited LH in Figure 2, for at least two reasons. First, a strong sensitivity of evapotranspiration to soil moisture (Figure 2) does not imply that the soil moisture variations are locally strong enough to induce large evapotranspiration variations and thus large impacts on air temperature (Figure 3c). Second, as noted previously, the plotted sensitivity also includes a signal of the size of the precipitation corrections, and so will be enhanced where the differences between the model-generated and observation-corrected precipitation are larger.

Figure 3c is consistent with previous studies identifying hot-spots of strong coupling between the land and T^{2m}. In particular Koster et al. (2006) and Miralles et al. (2012) both identify similar regions of strong coupling centered on the central US/central America and the Sahel, although they do not agree as well over south Asia. Over South Asia Koster et al. (2006) does not locate a hotspot, while Miralles et al. (2012) identifies India as having the strongest coupling, and Figure 3c suggests patchy regions of coverage spanning from southeast Asia through the north of India.

For reference, the corresponding maps for the Austral summer (December-January-February) are shown in supplemental Figure 1 for $R^2_{anom}(LH, SM)$ and supplemental Figure 2 for the sensitivity to the precipitation corrections. In supplemental Figure 1, the $R^2_{anom}(LH, SM)$ over Austral summer again shows the expected pattern of moisture-limited LH in drier areas of the summer hemisphere (almost everywhere, outside of the tropics). As with the Boreal summer, regions of moisture-limitation LH extend into the winter Hemisphere. However, the effect of reduced radiation close to the poles is now evident in the switch to energy-limited LH, even in arid regions that are poleward of around 50° (such as central Asia). Supplemental Figure 2 shows strong sensitivity of T^{2m}_{max} to the precipitation corrections across nearly all of the southern Hemisphere, including the Amazon and tropical Africa. Since these latter two areas typically have saturated soils, this strong signal is unlikely due to the precipitation-soil moisture pathway, and is perhaps due to sensitivity of evaporative cooling from the canopy interception to changes in precipitation supply to the interception reservoir.

c. Biases over Boreal summer

In Section 3.a, the biases in the reanalyses' global land energy budgets were provided as annual means. The seasonal cycle of the monthly mean global land biases (not shown) reveal that the largest global land biases for all budget terms occur in the Boreal summer (JJA). Below, maps of these JJA biases are presented and discussed, together with the corresponding biases in 2 m air temperatures.

1) ENERGY BUDGET TERMS

Figure 4 shows maps of the reanalyses' JJA biases in LH and SH compared to each of GLEAM and MTE. For LH, the regions of positive and negative biases relative to GLEAM or MTE are similar (compare the first and second columns of Figure 4). For both, the LH biases depend on the local LH regime, with energy-limited regions (low $R^2_{anom}(LH,SM)$ in Figure 2) generally having larger positive LH biases ($> 20\,Wm^{-2}$; e.g., for MERRA-2 in Figures 4d,e across the tropics, south Asia, and the northern high latitudes), while moisture-limited regions (high $R^2_{anom}(LH,SM)$ in

Figure 2) tend to have smaller biases (magnitude $<10 Wm^{-2}$). Consequently, the spatial correlation between $R^2_{anom}(LH,SM)$ (as plotted in Figure 2) and the MERRA-2 LH biases is -0.65 for GLEAM and -0.73 for MTE.

The MERRA LH biases (Figures 4j,k) show some of the same features as for MERRA-2, again with a tendency for large positive biases in energy-limited LH regimes. The most prominent difference is the sharp bias gradient in MERRA around 10°S (most notable in South America). As discussed in Section 2.b, this is associated with the unrealistically large rainfall interception reservoir in MERRA, combined with the MERRA precipitation errors; these problems have been alleviated in MERRA-2 (and MERRA-Land). Additionally, there are some isolated regions of large positive biases in moisture-limited regimes in MERRA that are removed in MERRA-2 (and MERRA-Land), such as in Mexico and south India.

Overall, in energy-limited regions ($R^2_{anom}(LH,SM)$ <0.5 in Figure 2) the area-averaged LH bias in MERRA-2 (25.5 Wm^{-2} compared to GLEAM, 29.9 Wm^{-2} compared to MTE) was slightly higher than for MERRA (24.1 Wm^{-2} compared to GLEAM, 27.6 Wm^{-2} compared to MTE), both of which are much higher than for MERRA-Land (11.3 Wm^{-2} compared to GLEAM, and 7.6 Wm^{-2} compared to MTE). In contrast, in moisture-limited LH regions ($R^2_{anom}(LH,SM)$ >0.5 in Figure 2), the area-averaged LH bias is highest in MERRA (7.0 Wm^{-2} compared to GLEAM, 5.2 Wm^{-2} compared to MTE), and reduced in MERRA-2 (3.8 Wm^{-2} compared to GLEAM, 1.5 Wm^{-2} compared to MTE), and even further reduced in MERRA-Land (0.3 Wm^{-2} compared to GLEAM, -2.9 Wm^{-2} compared to MTE).

The third column of Figure 4 shows the reanalyses biases in SH compared to MTE. In general, the SH biases for each reanalyses have an inverse relationship with the LH biases in the first two columns (for MERRA-2, the spatial correlation between the SH biases and the LH biases is -0.68 for GLEAM LH and -0.78 for MTE LH). Consequently, the evaporative fraction

(EF=LH/(LH+SH)) biases compared to MTE in the first column of Figure 5 show a spatial pattern very similar to that of the LH biases (for MERRA-2, the spatial correlation between MTE LH and EF biases is 0.83).

The sum of LH and SH approximates the net incoming radiation (after neglecting the ground heat flux and temporal change in T_{skin}). The second and third columns of Figure 5 show, respectively, the biases in the reanalyses LH+SH sum compared to MTE and the biases in their R_{net} compared to CERES-EBAF. There is a weak agreement between the R_{net} biases suggested by MTE and CERES-EBAF (for MERRA-2, the spatial correlation is 0.46). Comparison to MTE (Figures 5, second column) suggests that the reanalyses net surface radiation tends to be overestimated, with the largest biases ($>30\ Wm^{-2}$) occurring over the Amazon, the horn of Arica, and the Tibetan Plateau. While comparison to CERES-EBAF (Figure 5, third column) also suggests relatively large positive biases over the Tibetan Plateau and the horn of Africa, these positive biases are smaller in both magnitude and regional extent than was suggested by MTE. Additionally, CERES-EBAF also indicates strong negative biases ($<$-30 Wm^{-2}) over the Sahel and the southeast US, particularly in MERRA-Land (Figure 5i) and MERRA (Figure 5l). Finally, inter-comparing the R_{net} biases for each reanalyses shows qualitatively that the broad patterns are similar in MERRA-2 and MERRA (also MERRA-Land), although MERRA has a tendency towards larger (positive and negative) biases.

There is no obvious correspondence between the regional biases in the LH (compared to GLEAM or MTE) and the regional biases in R_{net} (compared to either MTE LH+SH or CERES-EBAF). For example, the spatial correlations are less than 0.1 between the MERRA-2 LH bias (implied by comparison to GLEAM or MTE), and the MERRA-2 LH+SH bias (implied by MTE). Likewise, the spatial correlations are again less than 0.1 between the MERRA-2 LH bias (implied by GLEAM of MTE) and the MERRA-2 R_{net} bias (implied by CERES-EBAF). This suggests then that the pattern of regional biases in the reanalyses LH for JJA (compared to either GLEAM or MTE) are associated with differences in the partitioning of incoming radiation into LH and SH, rather than with differences in the surface radiation (compared to MTE or CERES-EBAF) itself.

While radiation biases do not appear to be the main predictor of LH biases, biased radiation will results in biased LH and/or SH. Hence, we have partitioned the JJA R_{net} bias between MERRA-2 and CERES-EBAF into the individual contributions from each radiation term. Figure 6 shows the JJA biases between MERRA-2 and CERES-EBAF for the SW_{net}, LW_d, and LW_u. In terms of the direction of the biases, the broad patterns of regional biases in the radiation terms are unchanged from MERRA (not shown). The direction of the regional R_{net} biases for MERRA-2 in Figure 5f largely mirror the regional SW_{net} biases in Figure 6d (spatial correlation: 0.75), the main exception being over the southeast US. The LW biases are somewhat balanced, in that both are negative across most of the domain, with the LW_d bias in Figure 6e typically being slightly more negative than the LW_u bias in Figure 6f. Both have relatively large negative biases (magnitude $> 30\ Wm^{-2}$) in northern hemisphere desert regions, and smaller (magnitude: 10-20 Wm^{-2}) negative biases elsewhere. The spatial distribution of the SW_{net} biases mirrors that of the downwelling shortwave (SW_d, not shown), indicating that the SW_{net} biases are primarily driven by SW_d differences rather than differences in the surface albedo used in CERES-EBAF and GEOS-5. The above patterns of overestimated SW_{net} (or SW_d) and underestimated LW_d across much of the globe are consistent with a known tendency for the GEOS-5 systems to underestimate mid-latitude continental cloud cover (Molod et al. 2012; Wang and Dickinson 2013; Gelaro and Coauthors 2015).

The LW_u is calculated from the T_{skin}, and the negative biases in MERRA-2 (and also MERRA and MERRA-Land) indicate a cool bias in the model T_{skin}. At 285 K, a LW_u bias of 10 Wm^{-2} is roughly equivalent to a T_{skin} bias of 2 K. Recall that the CERES-EBAF LW_u is not independent of the MERRA suite of reanalyses, due to its use of GEOS-5 T_{skin}. However, the input GEOS-5 T_{skin}

is adjusted within the CERES-EBAF algorithm to constrain the TOA irradiance, so comparison of GEOS-5 and CERES LW_u indicates the adjustment required to the GEOS-5 T_{skin} to balance the TOA fluxes. Previous work has also suggested that the GEOS-5 T_{skin} is underestimated, particularly in dry regions. For example, in agreement with our Figure 6f, Draper et al. (2015) found large cool biases in the GEOS-5 T_{skin} over desert regions in summer (their Fig. 5), compared to remotely sensed observations. As argued in Draper et al. (2015), this GEOS-5 T_{skin} cool bias is, at least in part, caused by the model's T_{skin} definition differing from that of a true skin layer from which LW_u is emitted (or as is observed in the thermal infrared).

In summary, the pattern of regional LH biases in the reanalyses suggested by GLEAM and MTE are very similar. This result adds confidence to the use of GLEAM and MTE for estimating regional biases in the reanalyses. As with the annual global land averages in Figure 1, the maps presented here suggest that MERRA-2 and MERRA (but not MERRA-Land) have a general tendency to overestimate LH. If the GLEAM, MTE, and CERES-EBAF regional means are assumed to be more accurate than the reanalyses, the above comparisons suggest that in energy-limited regions, MERRA-2 (and MERRA) overestimate LH due to an overestimated evaporative fraction (i.e., too much incoming radiation is converted to LH rather than SH). There is little change in the global average biases from MERRA to MERRA-2. However, there are some isolated regions in Mexico and south Asia that are typified by moisture-limited LH, where MERRA has positive LH biases associated with overestimated EF, while MERRA-2 and MERRA-Land have much smaller biases. The precipitation corrections in MERRA-2 (and MERRA-Land) removed a relatively large amount of precipitation across these locations (Reichle et al. (2017b); their Figure 3b), strongly suggesting that the use of precipitation observations in these products reduced the LH biases.

2) Air temperature

The biases in the MERRA-2 and MERRA JJA monthly mean daily minimum, daily maximum, and diurnal range in T^{2m}, relative to the CRU data set, are shown in Figure 7 (T^{2m} is not calculated by the land-only MERRA-Land system). For the daily minimum T^{2m} (T^{2m}_{min}) in the left column, both reanalyses tend towards positive (warm) biases, particularly MERRA. For the daily maximum T^{2m} (T^{2m}_{max}) in the center column, MERRA-2 tends towards cool biases, with patches of warm biases across central Asia and the Arabian Peninsula (investigation of the large positive bias in the Arabian Peninsula suggests it is associated with an error in the CRU reference data, rather than the reanalyses). For MERRA, these patches of positive bias are expanded to cover most of the desert region in the northern hemisphere, and also much of the southern hemisphere. For the diurnal temperature range (DTR) in the third column, the MERRA-2 biases inherit the broad spatial pattern of the T^{2m}_{max} biases, while for MERRA some of the large positive T^{2m}_{max} biases are offset in the DTR by co-located positive T^{2m}_{min}.

The LH and SH biases in Figures 4 and the DTR biases in Figure 7 show some of the expected regional similarities. In particular, in the high latitudes and the Amazon MERRA-2 has relatively large positive LH biases (and negative SH biases) and relatively large negative DTR biases. MERRA also has overestimated LH and underestimated DTR in the same regions, as well as in southeast Asia and central America. This is consistent with an underestimated DTR caused by underestimated SH (and overestimated LH), particularly given that the R_{net} bias is generally neutral in these regions in Figure 5. It should however be noted that the high latitudes and the Amazon regions are both data-scarce, and both the reanalyses and reference data sets are less well constrained. In other regions there is less correspondence. For example the western US also has underestimated DTR for MERRA and MERRA-2, while neither GLEAM nor MTE suggests overestimated LH. Over all, the spatial correlations between the LH biases and DTR biases are rather low (for MERRA-2, they are -0.38 for GLEAM and -0.47 for MTE).

Recall that in Section 3.c.1 above, the CERES-EBAF comparison suggested that the MERRA-2 (and MERRA) T_{skin} is generally biased cool, with larger cool biases in desert areas. However, a comparison of the LW_u biases in Figure 6f to the T^{2m}_{min} and T^{2m}_{max} biases in Figures 7d,e shows little correspondence between them, and in particular the regions of relatively large cool T_{skin} biases (underestimated LW_u) in the northern hemisphere deserts do not have cool biases in either T^{2m}_{max} and T^{2m}_{min}. This apparent contradiction between the temperature biases suggested by comparison to the CERES-EBAF LW_u ($\sim T_{skin}$) and the CRU T^{2m}does not necessarily imply that one of these data sets is incorrect, given the likelihood mentioned above that the model T_{skin} biases are at least partly associated with the model definition of T_{skin}.

d. Turbulent heat flux anomaly correlations over Boreal summer

Here the monthly mean turbulent heat flux time series are evaluated over Boreal summer based on their temporal correlations (R_{anom}) with the reference data sets. Figure 8 shows maps of the JJA R_{anom} for each of the NASA reanalyses (MERRA-2, MERRA-Land, and MERRA) and ERA-Interim, with the R_{anom} calculated separately vs. each of the GLEAM and MTE turbulent heat fluxes. For LH, the regional patterns in the R_{anom} vs. either GLEAM (Figure 8, first column) or MTE (Figure 8, second column) show some similar features (for MERRA-2, spatial correlation between Figures 8a and 8b: 0.69). Comparison to Figure 2 again suggests some dependence on the LH regime. In the Northern Hemisphere, the LH R_{anom} is generally highest ($\sim$ 0.6) in regions where LH is moisture-limited, and generally much lower ($<$0.2) where LH is energy-limited. The two exceptions are the high latitudes, which have high LH R_{anom} and energy-limited LH, and the

Sahara, which has low LH R_{anom} and is moisture-limited (although LH variability in the Sahara is very low, making the signal susceptible to noise).

The R_{anom} patterns for ERA-Interim in the final row of Figure 8 provide some additional context for evaluating the NASA reanalyses. The LH R_{anom} values are generally higher for ERA-Interim than for the NASA reanalyses. As for MERRA-2, the ERA-Interim R_{anom} vs MTE is relatively low in many energy-limited LH regimes (including the eastern US, tropics, and south Asia), while the R_{anom} for ERA-Interim vs. GLEAM is more spatially consistent, in contrast to the R_{anom} for MERRA-2. The relatively high R_{anom} between GLEAM and ERA-Interim LH in energy-limited LH regimes may well be due to GLEAM having used ERA-Interim radiation and temperature, since it is in these regions that these fields will have the strongest influence on the LH. On the other hand, the lower R_{anom} between the NASA reanalyses and the LH reference data sets (and also between ERA-Interim and MTE) could be attributed to errors in both the reference data sets and the reanalyses under energy-limited conditions. For MTE, this result was expected because MTE is thought to be more reliable in estimating temporal variability in moisture limited areas, since its temporal variability is largely driven by fPAR (Jung et al. 2010).

Moving on to SH, the third column of Figure 8 shows the R_{anom} vs. MTE for each reanalysis. The regional patterns are similar to those for LH, with higher R_{anom} ($>$0.5) in moisture-limited LH regions, and lower ($<$ 0.2) values elsewhere. ERA-Interim R_{anom} vs. MTE is generally higher than the NASA reanalyses, with values greater than 0.5 across most of the globe (and particularly in the Northern Hemisphere). Despite the improved LH from MERRA-Land, the SH R_{anom} vs. MTE is lower than for MERRA (or MERRA-2).

Globally averaged, the rank order of the mean LH R_{anom}, while rather low, is the same vs. either GLEAM or MTE and follows the expected progression of improvement from MERRA, to MERRA-Land, and then to MERRA-2. GLEAM suggests a larger improvement, from a globally averaged R_{anom} of 0.39 for MERRA to 0.48 for MERRA-2, with MERRA-Land falling in between (0.45). MTE suggests an improvement from 0.29 for MERRA to 0.34 for MERRA-2, with MERRA-Land again falling in between (0.32). For SH, the globally averaged R_{anom} vs. MTE is similar for MERRA (0.36) and MERRA-2 (0.37), but is much lower for MERRA-Land (0.28). For ERA-Interim, the global mean R_{anom} for LH is ~0.1 higher than for MERRA-2 (0.60 vs. GLEAM, and 0.44 vs. MTE) and ~0.2 higher for SH (0.46 vs. MTE). The better agreement between ERA-Interim and the reference data sets could be a consequence of the land surface updates applied in ERA-Interim, which indirectly targets the turbulent heat fluxes. (Although recall that the relatively strong agreement between the GLEAM and ERA-Interim LH will partly reflect their dependence; see Section 2.c.2).

e. Comparison to Fluxnet tower data

Since the reference data sets used above do not represent direct observations, we now compare the globally-averaged LH and SH statistics from Section 3.a (for the annual mean turbulent heat fluxes over land), and Section 3.d (for the mean JJA R_{anom}) to statistics calculated against Fluxnet-2015 tower observations. Figure 9 shows the annual mean of the turbulent fluxes averaged across the 20 tower sites for the Fluxnet (eddy-covariance) measurements themselves and for each reanalysis and reference data set averaged across the 20 Fluxnet locations, with the global land annual means (from Figure 1) included for reference. For LH, comparison to the Fluxnet observations agrees with the results from the global land comparison in Section 3.a, again suggesting that the MERRA-2 LH is biased high, although the Fluxnet observations suggest a larger bias (of 12 Wm^{-2}, or 30%) than was suggested by the global comparison (estimated as 6 Wm^{-2} in Section 3.a). Averaged across the 20 Fluxnet sites, the MTE LH is very close to the Fluxnet data (within 0.5 Wm^{-2}), while GLEAM is slightly higher. For the interested reader, supplemental

Figure 3 shows scatterplots comparing the MERRA-2 and reference data set LH annual means at the 20 individual sites.

For SH, the Fluxnet observations agree less well with the global land comparison. First, the annual mean of the Fluxnet data is about 10 Wm^{-2} below the global mean estimates from the other reference data sets. For each of the global reference data sets and reanalyses, the annual average over the 20 Fluxnet sites is also 15-20 Wm^{-2} lower than the global average, suggesting that the relatively low Fluxnet annual mean is associated with the spatial sampling of the Fluxnet sites. Second, averaged across the Fluxnet sites, the Fluxnet mean SH is close to that of MERRA-Land, and above that of MERRA-2 (by 6 Wm^{-2}, 18 %). In contrast, for the global averages in Section 3.a the reference data sets were all close to MERRA-2 (and MERRA), with MERRA-Land standing out as being biased high.

Figure 10 shows the JJA R_{anom} averaged over the 20 Fluxnet sites for each reanalyses vs. each of Fluxnet, GLEAM, and MTE, with the global average JJA R_{anom} from Section 3.d also included for GLEAM and MTE. The R_{anom} for the Fluxnet data are quite low, which is somewhat expected due to the mismatch in spatial representation between the tower-based observations and the reanalysis. Nonetheless, the Fluxnet R_{anom} (as well as the GLEAM and MTE R_{anom} at the same locations) indicates similar relative reanalysis performance as the global mean R_{anom}. In particular, for LH MERRA-2 and MERRA-Land outperform MERRA, as also indicated by the global means. However, the one discrepancy is that the R_{anom} vs. the Fluxnet data is similar for ERA-Interim and MERRA-2, while the global comparisons (and also the GLEAM and MTE data averaged across the Fluxnet sites) all suggest that ERA-Interim outperforms MERRA-2 (giving mean R_{anom} around 0.1 higher). For SH, the rank order between the average JJA R_{anom} is the same from the Fluxnet data than from the global reference data sets, with the MERRA-Land R_{anom} again being lower than that for MERRA (and MERRA-2), and the ERA-Interim average R_{anom} being higher than that for MERRA-2.

It is notable that over the Fluxnet tower sites, both GLEAM and MTE have higher average R_{anom} with the reanalyses than the Fluxnet observations do. In particular, MTE was trained on an earlier generation of the Fluxnet data, and the higher mean R_{anom} vs. MTE than vs. Fluxnet suggests that the MTE algorithm has added coarse-scale information (similar quality control was applied here as was applied to the tower observations used in MTE). For the interested reader, supplemental Figure 4 shows scatterplots of the MERRA-2 LH R_{anom} vs. each reference data set at the 20 individual sites.

Note that for Fluxnet, the R_{anom} for (LH+SH), plotted in Figure 10c is consistently about 0.1 higher than the R_{anom} for either LH or SH separately. Decker et al. (2012) obtained a similar result for the correlation between reanalyses and tower observations. This indicates that the eddy covariance measurements and the reanalyses have a stronger agreement in the implied incoming radiation than in the partitioning of that radiation into LH and SH (this result is unchanged if the R_{anom} are calculated from the Fluxnet data that have not been energy balance-corrected). This could be a signal of errors in the partitioning within the reanalyses, or perhaps just as likely, this difference is associated with the spatial representation of the tower observations, since the incoming radiation is more spatially homogeneous than either LH or SH on its own.

f. Precipitation Corrections and Air Temperature Performance

Finally, we seek to establish whether the precipitation corrections in MERRA-2 influenced the local T^{2m}_{max}. We do this by comparing the performance of the MERRA-2 and MERRA T^{2m}_{max} to Figure 3c, which shows the MERRA-2 sensitivity to observed precipitation. Figure 11 shows the T^{2m}_{max} R_{anom} vs. CRU observations over JJA for MERRA-2 and MERRA. In general, the MERRA-

2 R_{anom} is high (> 0.7) across most of the domain, particularly in the high latitudes, with much lower (< 0.4) values across much of the tropics and parts of South America, Africa, and south Asia. Note that the latter regions all have relatively sparsely distributed CRU station data, which is likely contributing to the lower agreement with the reanalyses. Compared to MERRA, the greatest improvements in the MERRA-2 T_{max}^{2m} R_{anom} occurred in the eastern US, much of tropical South America and Africa, the Sahel, and parts of south Asia and China. There are also several regions where the T_{max}^{2m} R_{anom} is reduced, including northern South America, and much of southeast Asia. Overall, the global averaged T_{max}^{2m} R_{anom} vs. CRU was increased from 0.69 for MERRA to 0.75 for MERRA-2.

Comparing Figure 11c to Figure 3c, the regions with the strongest sensitivity of T_{max}^{2m} to the precipitation corrections generally have relatively large changes in the T_{max}^{2m} R_{anom} (including the Sahel, parts of south Asia, and central America). Consequently, where the metric in Figure 3c is above 0.25 (i.e., the observation-corrected precipitation explains at least 25% more of the MERRA-2 T_{max}^{2m} variance than the model-generated precipitation does), the area-averaged absolute change in the R_{anom} is 0.15, compared to an area-average absolute change of 0.07 elsewhere. This tendency toward relatively large change in the T_{max}^{2m} R_{anom} where T_{max}^{2m} is sensitive to the precipitation corrections suggests that the observed precipitation in MERRA-2 contributed to the change in T_{max}^{2m} performance. Additionally, the change in T_{max}^{2m} R_{anom} in these regions is generally, although not always, positive (giving an area averaged change in the R_{anom} of 0.06 where the metric in Figure 3c is greater than 0.25). In some of the instances where the T_{max}^{2m} R_{anom} is degraded, this can be traced back to errors in the precipitation observation data sets input into MERRA-2. For example, over Myanmar, the T_{max}^{2m} R_{anom} is decreased by more than 0.15, likely due to persistent local errors in the precipitation observations input into MERRA-2 (Reichle et al. 2017b). Finally, there are also regions with large changes in the T_{max}^{2m} R_{anom} outside of the regions of T_{max}^{2m} sensitivity to precipitation (the eastern US, tropical Africa and South America, and central China). The T^{2m}_{max} R_{anom} is increased in MERRA-2 across most of these regions, likely due to other advances (beyond the use of observed precipitation) in the MERRA-2 modeling and assimilation system.

4. Summary and conclusions

The land surface energy budgets of three reanalyses from NASA (MERRA, MERRA-Land, and MERRA-2) are compared here to the best available estimates from the literature and to (largely) independent global reference data sets. In terms of the global land annual averages, the results suggest that the MERRA-2 LH and SH are biased high by 5 Wm^{-2} and 6 Wm^{-2}, respectively, while SW_u has a large positive bias of 14 Wm^{-2}, SW_d is biased high by 3 Wm^{-2}, and the upwelling and downwelling LW components are biased low, by 11 Wm^{-2} and 13 Wm^{-2}, respectively. Compared to MERRA, this is a slight ($\sim$ 2 Wm^{-2}) reduction in the LH and SW_{net} biases, while the difference is even smaller for the LW terms ($\sim$ 1 Wm^{-2}). The radiation biases are associated with known issues in the GEOS-5 models used in the reanalyses, specifically a tendency to underestimate mid-latitude continental clouds (Wang and Dickinson 2013) and a cool bias in the model T_{skin} (Draper et al. 2015).

Compared to reference flux estimates from GLEAM and MTE over the Boreal summer (when both the fluxes themselves and their biases are greatest), the largest MERRA-2 LH biases (>20 Wm^{-2}, vs. either GLEAM or MTE) occur in regions where LH is energy-limited, such as in the high latitudes, the tropics, parts of south Asia, and the eastern US. The MERRA-2 LH biases are typically smaller in regions where LH is moisture-limited, which include the drier regions of the mid and low latitudes. In some of these moisture-limited regions (parts of south Asia and Mexico) the high bias in the MERRA LH was largely removed in MERRA-2 (and MERRA-Land), likely because the observed precipitation used in the latter was lower than that produced by the MERRA (or MERRA-2) modeling systems. Finally, comparison to the evaporative fraction from MTE and to R_{net} from CERES-EBAF or as inferred from MTE LH+SH indicates that the regional biases in the reanalyses LH are generally associated with differences in the partitioning of R_{net} into LH and SH rather than with differences in the radiation input.

The temporal agreement between the reanalyses and the reference data sets over Boreal summer was measured using the monthly anomaly correlation (R_{anom}) over JJA. For LH, the R_{anom} between the reanalyses and the reference data sets (GLEAM and MTE) again showed some dependency on the LH regime, with a tendency towards better agreement where LH is moisture-limited than where it is energy-limited. The lower agreement in energy-limited regions does not necessarily imply poorer performance in the reanalyses, as it may be due to errors in the reference data sets. The globally averaged R_{anom} values show the expected improvement in skill with each new NASA reanalyses. For example, MERRA-2 has slightly better globally averaged LH R_{anom} (0.48 vs GLEAM) than MERRA-Land (0.45), which is substantially better than MERRA (0.39). The R_{anom} was also calculated for the monthly mean daily T_{max}^{2m} vs. CRU reference data over JJA. Averaged over global land, the JJA T_{max}^{2m} R_{anom} vs. CRU increased from 0.69 for MERRA to 0.75 for MERRA-2. The results presented above for the regional biases and R_{anom} were based on the Boreal summer, however the same analysis has been performed over the Austral summer (not shown), yielding qualitatively similar results.

The use of observed precipitation in MERRA-2 was motivated by the hope that the subsequent improvements in simulated soil moisture would lead to the improved partitioning of incoming radiation between latent and sensible heating, ultimately leading to improvements in the diurnal evolution of the boundary layer. It is difficult, however, to unequivocally attribute the improvements in MERRA-2 to the use of observed precipitation because MERRA-2 includes many other modeling and assimilation advances relative to MERRA. Nonetheless, many of the improvements in the MERRA-2 LH and T^{2m} are consistent with the changes expected from the use of observed precipitation. MERRA-2 and MERRA-Land have smaller positive LH biases and higher LH R_{anom} than MERRA in regions where LH is moisture-limited and thus sensitive to precipitation (south Asia and the western US). This is most easily explained by the forcing of the land surface with observed precipitation in MERRA-2. Additionally, regions where the MERRA-2 JJA T^{2m}_{max} was most sensitive to the precipitation corrections (the Sahel, central US, and parts of south Asia), generally experience larger changes in the T^{2m}_{max} R_{anom} from MERRA to MERRA-2. However, the changes in R_{anom} in these areas are not uniformly positive, and in some cases degraded T^{2m}_{max} R_{anom} can be traced back to problems in the input precipitation data sets (e.g., over Myanmar). In the future, the use of precipitation corrections could be enhanced by also implementing a land data assimilation scheme to update the model soil moisture according to observations (e.g., Draper et al. (2011); Dharssi et al. (2011); De Lannoy and Reichle (2016)). By making use of remotely sensed observations, the land data assimilation would be particularly valuable in regions where the rain-gauge network is sparse or has known problems (e.g., in Africa and parts of southeast Asia).

However, some of the largest biases and lowest R_{anom} for the MERRA-2 turbulent fluxes occur where the LH is energy-limited and thus less sensitive to improvements in the precipitation and soil moisture. Hence, future efforts to improve the MERRA-2 land surface turbulent fluxes would best be focused on other facets of the modeling and assimilation. Specifically, future GEOS-5 development should focus on the overestimated evaporative fraction where LH is energy-limited. Additionally, even though the MERRA-2 R_{net} is relatively unbiased (compared to CERES-EBAF), there are large compensating biases in the individual SW and LW radiation fluxes that are 2-3 times the magnitude of the LH biases in terms of the global land annual averages. Reducing the cloud bias in the atmospheric model will help these biases, as will re-defining the model T_{skin} to generate a LW_u more consistent with observations.

Finally, the SH results for MERRA-Land are troubling. While MERRA-Land did have the desired reduction in the LH biases compared to MERRA (to 1 Wm^{-2} in the global land annual average), it also had a compensating, and much larger, increase in the SH bias (up to 15 Wm^{-2} in the global land average). Additionally, the JJA R_{anom} compared to MTE were reduced from MERRA to MERRA-Land (from a global average of 0.36 to 0.28), despite the LH R_{anom} being increased. The cause of the degraded SH in MERRA-Land is presently unknown, but given the otherwise similar MERRA and MERRA-Land land surface models and meteorological forcing, an obvious possibility is that the use of observed precipitation in an offline (land-only) replay of an analysis, such as MERRA-Land, can lead to inconsistencies in the forcing (e.g., warm and dry air, stemming from dry conditions in MERRA, overlying cold ground induced by high antecedent rainfall from the observations). Such inconsistencies would not appear in MERRA or (as much) in MERRA-2, given the coupling in the reanalyses of the land surface state with the overlying atmosphere.

While this work focused on evaluating surface energy fluxes in MERRA-2, the findings have relevance to anyone interested in designing a methodology to evaluate global estimates of turbulent heat fluxes. The gridded LH reference data sets (GLEAM and MTE) had better agreement with the reanalyses time series (as measured by R_{anom}), and were more useful for evaluating the reanalysis output than were the tower observations. In particular they offer (near-) global coverage across several decades, at similarly course resolution to the reanalyses. In the absence of a recognized truth for LH (or other similar terms), the recommended evaluation strategy is to compare the product under evaluation to multiple data sets. However, given the uncertainty in the available reference data sets, extra care is necessary to understand the methodology, input data, assumptions, and potential dependencies and weaknesses of each reference data set. This process relies on expert judgement and inevitably introduces some subjectivity into the interpretation of the results. Further development of global gridded LH data sets (including the quality and quantity of ground-'truth' observations), to increase their confidence would obviously be of great benefit to this process.

The GLEAM and MTE reference data sets used here are independent of each other and are based on very different methodologies, thus providing complementary information for use in an evaluation. However, given the use of the common precipitation input data in GLEAM as in MERRA-2, and the fact that MTE data is not optimized to estimate interannual variability, LH estimates from a third reference data set would be useful. Emerging global and multi-decadal land surface flux data sets based on an energy balance approach (Anderson et al. 2011), or alternative observational frameworks (Alemohammad et al. 2017) would provide useful complements to GLEAM and MTE for a more comprehensive analysis.

Acknowledgments. Funding for this work was provided by the NASA Modeling, Analysis, and Prediction program. Computational resources were provided by the NASA High-End Computing Program through the NASA Center for Climate Simulation. The authors acknowledge the teams that produce and publish the GLEAM, MTE, CERES-EBAF, ERA-Interim, MERRA, MERRA-Land, and MERRA-2 products. Additionally, we are grateful to Diego Miralles (VU University Amsterdam/Ghent University), Martin Jung (Max Planck Institute for Biogeochemistry), and Seiji Kato (NASA Langley Research Center) for their thoughtful feedback on this work, and detailed advice on the use of GLEAM, MTE, and CERES-EBAF, respectively. The FLUXNET eddy covariance data processing and harmonization was carried out by the European Fluxes Database Cluster, AmeriFlux Management Project, and Fluxdata project of FLUXNET, with the support of CDIAC and ICOS Ecosystem Thematic Center, and the OzFlux, ChinaFlux and AsiaFlux offices.

References

Alemohammad, S. H., and Coauthors, 2016: Water, Energy, and Carbon with Artificial Neural Networks (WECANN): A statistically-based estimate of global surface turbulent fluxes using solar-induced fluorescence. *Biogeosciences*, **15**, 4101–4124, doi:10.5194/bg-14-4101-2017.

Anderson, M., and Coauthors, 2011: Mapping daily evapotranspiration at field to continental scales using geostationary and polar orbiting satellite imagery. *Hydrology and Earth System Sciences*, **15**, 223–239, doi:10.5194/hess-15-223-2011.

Baldocchi, D., and Coauthors, 2001: FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*, **82**, 2415–2434, doi:10.1175/1520-0477(2001)082⟨2415:FANTTS⟩2.3.CO;2.

Beck, H., A. van Dijk, V. Levizzani, J. Schellekens, D. Miralles, B. Martens, and A. de Roo, 2017: MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences*, **21**, 589–615, doi:10.5194/hess-21-589-2017.

Betts, A., A. Tawfik, and R. Desjardins, 2017: Revisiting hydrometeorology using cloud and climate observations. *Journal of Hydrometeorology*, **18**, 939–955, doi:10.1175/JHM-D-16-0203.1.

CERES-EBAF, 2017: CERES_EBAF-Surface_Ed4.0, Data Quality Summary (May 26, 2017). Accessed Jun. 20, 2017 pp., https://ceres.larc.nasa.gov/documents/DQ_summaries/CERES_EBAF-Surface_Ed4.0_DQS.pdf.

Chen, M., W. Shi, P. Xie, V. B. S. Silva, V. E. Kousky, R. Wayne Higgins, and J. E. Janowiak, 2008: Assessing objective techniques for gauge-based analyses of global daily precipitation. *Journal of Geophysical Research: Atmospheres*, **113**, D04 110, doi:10.1029/2007JD009132.

De Lannoy, G., and R. Reichle, 2016: Assimilation of SMOS brightness temperatures or soil moisture retrievals into a land surface model. *Hydrology and Earth System Sciences*, **20**, 4895–4911, doi:10.5194/hess-20-4895-2016.

de Rosnay, P., G. Balsamo, C. Albergel, J. Munoz-Sabater, and L. Isaksen, 2014: Initialisation of land surface variables for numerical weather prediction. *Surveys in Geophysics*, **35**, 607–621, doi:10.1007/s10712-012-9207-x.

Decker, M., M. Brunke, Z. Wang, K. Sakaguchi, X. Zeng, and M. Bosilovich, 2012: Evaluation of the Reanalysis Products from GSFC, NCEP, and ECMWF Using Flux Tower Observations. *Journal of Climate*, **25**, 1916–1944, doi:10.1175/JCLI-D-11-00004.1.

Dee, D., and Coauthors, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137**, 553–597, doi:10.1002/qj.828.

Dharssi, I., K. Bovis, B. Macpherson, and C. Jones, 2011: Operational assimilation of ASCAT surface soil wetness at the Met Office. *Hydrology and Earth System Sciences*, **15**, 2729–2746, doi:0.5194/hess-15-2729-2011.

Draper, C., J.-F. Mahfouf, and J. Walker, 2011: Root-zone soil moisture from the assimilation of screen-level variables and remotely sensed soil moisture. *Journal of Geophysical Research*, **116**, D02 127, doi:10.1029/2010JD013829.

Draper, C., R. Reichle, and B. De Lannoy, G.and Scarino, 2015: A Dynamic Approach to Addressing Observation-Minus-Forecast Bias in a Land Surface Skin Temperature Data Assimilation System. *Journal of Hydrometeorology*, **16**, 449–464, doi:10.1175/JHM-D-14-0087.1.

Fluxnet, 2015: FLUXNET2015 Dataset. Accessed Aug 9, 2016 pp., http://fluxnet.fluxdata.org/data/fluxnet2015-dataset/.

Gelaro, R., and Coauthors, 2015: Evaluation of the 7-km GEOS-5 Nature Run. 285pp pp., NASA Technical Report Series on Global Modeling and Data Assimilation, NASA/TM-2014-104606, Vol. 36.

Gelaro, R., and Coauthors, 2017: The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, **30**, 5419–5454, doi:10.1175/JCLI-D-16-0758.1.

Global Modeling and Assimilation Office, 2008a: tavg1_2d_slv_Nx: MERRA 2D IAU Diagnostic, Single Level Meteorology, Time Average 1-hourly V5.2.0. Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD, Accessed Oct 1, 2016 pp., doi:10.5067/B6DQZQLSFDLH.

Global Modeling and Assimilation Office, 2008b: tavgM_2d_lnd_Nx: MERRA 2D IAU Diagnostic, Land Only States and Diagnostics, Monthly Mean V5.2.0. Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD, Accessed Oct 1, 2016 pp., doi:10.5067/XOHTIIK0W9RK.

Global Modeling and Assimilation Office, 2008c: tavgM_2d_mld_Nx: MERRA Simulated 2D Incremental Analysis Update (IAU) MERRA-Land reanalysis, GEOSldas-MERRALand, Time

Average Monthly Mean V5.2.0. Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD, Accessed Oct 1, 2016 pp., doi:10.5067/K9PCGOMQ1XP1.

Global Modeling and Assimilation Office, 2015a: MERRA-2 tavg1_2d_lfo_Nx: 2D, 1-Hourly, Time-Averaged, Single-Level, Assimilation, Land Surface Forcings, V5.12.4. Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD, Accessed Oct 1, 2016 pp., doi:10.5067/L0T5GEG1NYFA.

Global Modeling and Assimilation Office, 2015b: MERRA-2 tavgM_2d_lnd_Nx: 2D, Monthly Mean, Time-Averaged, Single-Level, Assimilation, Land Surface Diagnostics, V5.12.4. Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD, Accessed Oct 1, 2016 pp., doi:10.5067/8S35XF81C28F.

Global Modeling and Assimilation Office, 2015c: MERRA-2 tavgM_2d_slv_Nx: 2D, Monthly Mean, Time-Averaged, Single-Level, Assimilation, Single-Level Diagnostics, V5.12.4. Goddard Earth Sciences Data and Information Services Center (GES DISC), Greenbelt, MD, Accessed Oct 1, 2016 pp., doi:10.5067/AP1B0BA5PD2K.

Harris, I., P. Jones, T. Osborn, and D. Lister, 2014: Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset. *International Journal of Climatology*, **34**, 623642, doi:10.1002/joc.3711.

Huffman, G., R. Adler, D. Bolvin, and G. Gu, 2009: Improving the global precipitation record: GPCP Version 2.1. *Geophysical Research Letters*, **36 (17)**, L17 808, doi:10.1029/2009GL040000.

Jiménez, C., and Coauthors, 2011: Global intercomparison of 12 land surface heat flux estimates. *Journal of Geophysical Research: Atmospheres*, **116**, D02 102, doi:10.1029/2010JD014545.

Jung, M., M. Reichstein, and A. Bondeau, 2009: Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model. *Biogeosciences*, **6**, 2001–2013, doi:10.5194/bg-6-2001-2009.

Jung, M., and Coauthors, 2010: Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature*, **467**, 951–954, doi:10.1038/nature09396.

Jung, M., and Coauthors, 2011: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research - Biogeosciences*, **116**, G00j07, doi: 10.1029/2010jg001566.

Kato, S., N. Loeb, F. Rose, D. Doelling, D. Rutan, T. Caldwell, L. Yu, and R. Weller, 2013: Surface Irradiances Consistent with CERES-Derived Top-of-Atmosphere Shortwave and Longwave Irradiances. *Journal of Climate*, **26**, 2719–2740, doi:10.1175/JCLI-D-12-00436.1.

Koster, R., G. Salvucci, A. Rigden, M. Jung, G. Collatz, and S. Schubert, 2015: The pattern across the continental United States of evapotranspiration variability associated with water availability. *Frontiers in Earth Science*, **3**, 35pp, doi:10.3389/feart.2015.00035.

Koster, R., and Coauthors, 2006: GLACE: The Global Land-Atmosphere Coupling Experiment. Part I: Overview. *Journal of Hydrometeorology*, **7**, 590–610, doi:10.1175/JHM510.1.

L'Ecuyer, T., and Coauthors, 2015: The observed state of the Energy budget in the early 21st Century. *Journal of Climate*, **28**, 8319–8346, doi:10.1175/JCLI-D-14-00556.1.

Martens, B., and Coauthors, 2017: GLEAM v3: satellite-based land evaporation and root-zone soil moisture. *Geoscientific Model Development*, **10**, 1903–1925, doi:10.5194/gmd-10-1903-2017.

Miralles, D., T. Holmes, R. de Jeu, J. Gash, A. Meesters, and A. Dolman, 2011: Global land-surface evaporation estimated from satellite-based observations. *Hydrology and Earth System Sciences*, **15**, 453–469, doi:10.5194/hess-15-453-2011.

Miralles, D., M. van den Berg, A. Teuling, and R. de Jeu, 2012: Soil moisture-temperature coupling: A multiscale observational analysis. *Geophysical Research Letters*, **39**, L21 707, doi: 10.1029/2012GL053703.

Molod, A., L. Takacs, M. Suarez, and J. Bacmeister, 2015: Development of the GEOS-5 atmospheric general circulation model: evolution from MERRA to MERRA-2. *Geoscientific Model Development*, **8**, 1339–1356, doi:10.5194/gmd-8-1339-2015.

Molod, A., L. Takacs, M. Suarez, J. Bacmeister, I.-S. Song, and A. Eichmann, 2012: The GEOS-5 atmospheric general circulation model: Mean climate and development from MERRA to Fortuna . 117pp pp., NASA Technical Report Series on Global Modeling and Data Assimilation, NASA/TM-2014-104606, Vol. 28.

Mueller, B., and Coauthors, 2011: Evaluation of global observations-based evapotranspiration datasets and IPCC AR4 simulations. *Geophysical Research Letters*, **38**, L06 402, doi:10.1029/2010GL046230.

Mueller, B., and Coauthors, 2013: Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis. *Hydrology and Earth System Sciences*, **17**, 3707–3720, doi: 10.5194/hess-17-3707-2013.

NSIT, 2007: A NASA Earth science implementation plan for energy and water cycle research: Predicting energy and water cycle consequences of Earth system variability and change. 89pp pp., http://news.cisc.gmu.edu/doc/NEWS_implementation.pdf.

Reichle, R., C. Draper, Q. Liu, M. Girotto, S. Mahanama, R. Koster, and G. D. Lannoy, 2017a: Assessment of MERRA-2 land surface hydrology estimates. *Journal of Climate*, **30**, 2937–2960, doi:10.1175/JCLI-D-16-0720.1.

Reichle, R., R. Koster, G. De Lannoy, B. Forman, Q. Liu, S. Mahanama, and A. Toure, 2011: Assessment and enhancement of MERRA land surface hydrology estimates. *Journal of Climate*, **24**, 6322–6338, doi:10.1175/JCLI-D-10-05033.1.

Reichle, R., and Q. Liu, 2014: Observation-corrected precipitation estimates in GEOS-5. 18pp pp., NASA Technical Report Series on Global Modeling and Data Assimilation, NASA/TM-2014-104606, Vol. 35.

Reichle, R., Q. Liu, R. Koster, C. Draper, S. Mahanama, and G. Partyka, 2017b: Land surface precipitation in MERRA-2. *Journal of Climate*, **30**, 1643–1664, doi:10.1175/JCLI-D-16-0570.1.

Rienecker, M., and Coauthors, 2011: MERRA - NASA's Modern-Era Retrospective Analysis for Research and Applications. *Journal of Climate*, **24**, 3624–3648, doi:10.1175/JCLI-D-11-00015.1.

Schlosser, C., and X. Gao, 2010: Assessing evapotranspiration estimates from the Second Global Soil Wetness Project (GSWP-2) simulations. *Journal of Hydrometeorology*, **11**, 880–897, doi:10.1175/2010JHM1203.1.

Trenberth, K., J. Fasullo, and J. Kiehl, 2009: Earth's global energy budget. *Bulletin of the American Meteorological Society*, **90**, 311–323, doi:10.1175/2008BAMS2634.1.

University of East Anglia Climatic Research Unit, Harris I, and Jones, P, 2014: CRU TS3.22: Climatic Research Unit (CRU) Time-Series (TS) Version 3.22 of High Resolution Gridded Data of Month-by-month Variation in Climate (Jan. 1901- Dec. 2013). NCAS British Atmospheric Data Centre, doi:10.5285/18BE23F8-D252-482D-8AF9-5D6A2D40990C.

Wang, K., and R. Dickinson, 2013: Global atmospheric downward longwave radiation at the surface from ground-based observations, satellite retrievals, and reanalyses. *Reviews of Geophysics*, **51**, 150–185, doi:10.1002/rog.20009.

Wild, M., D. Folini, and M. Hakuba, 2015: The energy balance over land and oceans: an assessment based on direct observations and CMIP5 climate models. *Clim Dyn*, **44**, 3393–3429, doi:10.1007/s00382-014-2430-z.

Wilson, K., and Coauthors, 2002: Energy balance closure at FLUXNET sites. *Agricultural and Forest Meteorology*, **113**, 223–243, doi:10.1016/S0168-1923(02)00109-0.

Xie, P., and P. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bulletin of the American Meteorological Society*, **78**, 2539–2558, doi:10.1175/1520-0477(1997)078⟨2539:GPAYMA⟩2.0.CO;2.

LIST OF TABLES

TABLE 1. The reanalyses

Data set	Variables used	Output coverage and resolution (variable data set citation, where available)
MERRA-2		1980-ongoing, hourly, 5/8° x 0.5°
	LH,SH, LW_{net}, SW_{net}	global land (Global Modeling and Assimilation Office 2015b)
	LW_d	global surface (Global Modeling and Assimilation Office 2015a)
	T^{2m}_{max}, T^{2m}_{min}	global surface (Global Modeling and Assimilation Office 2015c)
MERRA-Land		1980-2016, hourly, 2/3° x 0.5°
	LH, SH, LW_{net}	global land (Global Modeling and Assimilation Office 2008c)
MERRA		1979-2015, hourly, 2/3° x 0.5°
	LH,SH, LW_{net}, SW_{net}	global land (Global Modeling and Assimilation Office 2008b)
	LW_d	global surface (-)
	T^{2m}_{max}, T^{2m}_{min}	global surface (Global Modeling and Assimilation Office 2008a)
ERA-Interim		1979 - ongoing, monthly mean, 79 km
	LH, SH	global surface

TABLE 2. The gridded reference data sets.

Data set	Variables used	Output coverage and resolution	Dependencies, error estimates where available
GLEAM v3.1a	LH	1980-2016, daily mean, 0.25° global land	Uses a precipitation data set that includes CPCU (used in MERRA-2, MERRA-Land) and ERA-Interim precipitation, uses T^{2m} and radiation from ERA-Interim. *c.f.* tower obs., average ubRMSE: 20 Wm^{-2}, average R_{anom}: 0.42. Full details: Section 2.c.1.
MTE	LH, SH	1982-2011 monthly mean, 0.5° global land, excluding non-vegetated regions	Trained on an earlier generation of the Fluxnet-2015 data set. Uses a CRU-based T^{2m} data set, and CPCU precipitation (neither strongly influences temporal behavior). *c.f.* withheld tower obs., average RMSE: 15 Wm^{-2}(LH & SH), average R_{anom} 0.57 (LH), 0.60 (SH). Full details: Section 2.c.2.
CRU v4.00	T^{2m}_{min}, T^{2m}_{max}	1901-2015 monthly means 0.5° global land (data not informed by station obs. have been removed)	Input station obs. will overlap with T^{2m} assimilated into ERA-Interim. Locally, will be more uncertain where input station obs. are sparse. Full details: Section 2.c.3.
CERES-EBAF, vn 4.0	SW_d, SW_u, LW_d, LW_u	Mar. 2000-Feb. 2016 monthly mean, 1° global surface	Uses atmospheric profile and T_{skin} from same system as used in the NASA re-analyses (results in strong dependence for LW_u, LW_d). *c.f.* ground obs. average RMSE: 12 Wm^{-2}(SW_d), 10 Wm^{-2} (LW_d). Full details: Section 2.c.4.

TABLE 3. Global annual land average energy budget from the NASA reanalyses (Wm^{-2}), estimated over an area of 130.2 million km^2.

	SW_d	SW_u	LW_d	LW_u	R_{net}	LH	SH
MERRA-2	204.6	40.7	312.6	385.5	91.0	47.8	42.2
MERRA-Land	as for MERRA			384.1	95.1	42.5	52.1
MERRA	206.5	40.9	313.7	386.7	92.6	50.4	41.2

LIST OF FIGURES

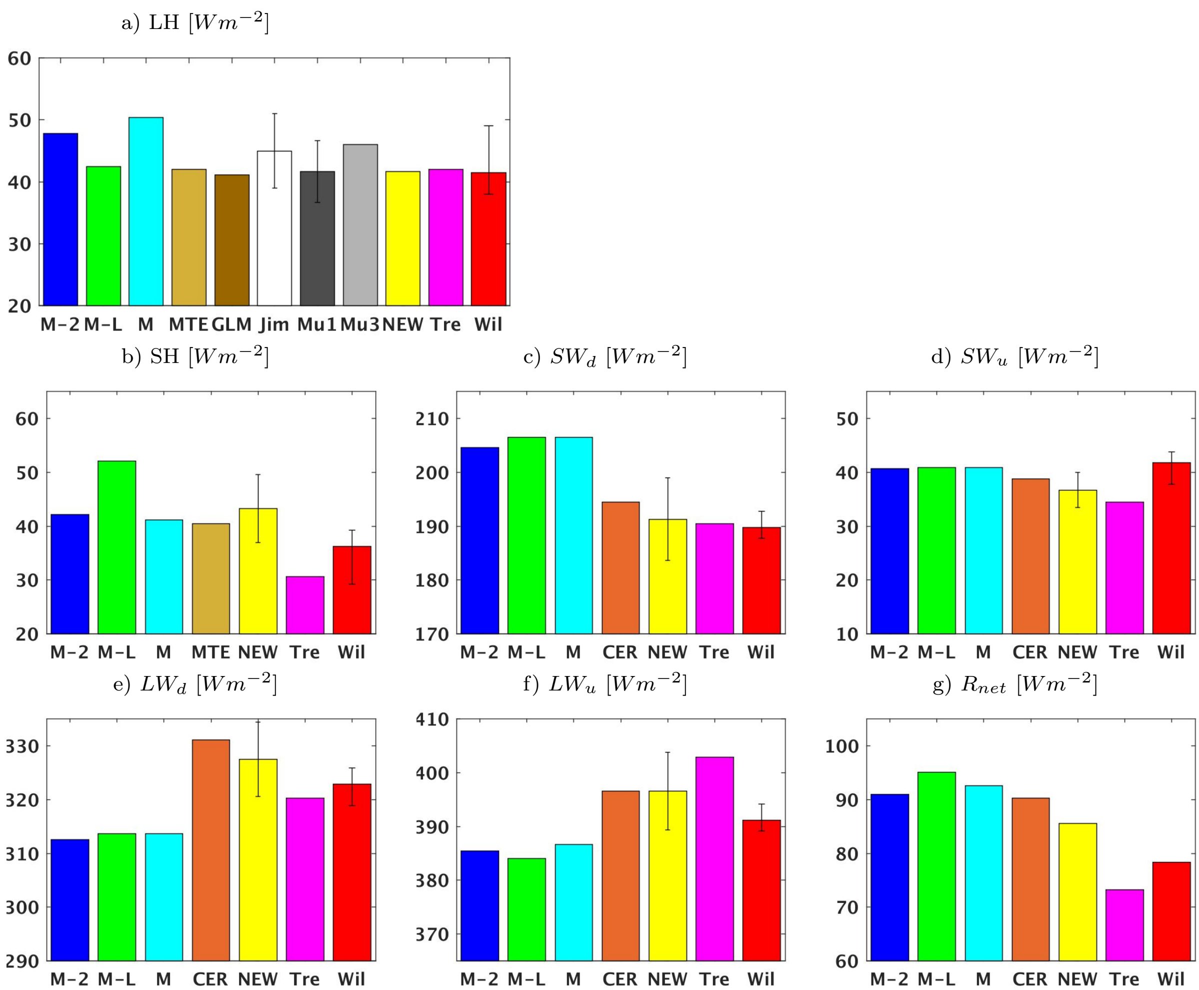

FIG. 1. The global annual mean energy budget over land from the reanalyses (MERRA-2 (M-2); MERRA-Land (M-L); MERRA (M)), the literature (NEWS (NEW), Trenberth et al. (2009) (Tre), Wild et al. (2015) (Wil), Jiménez et al. (2011) (Jim), Mueller et al. (2011) (Mu1), and Mueller et al. (2013) (Mu3)), and the gridded reference data sets (MTE, GLEAM (GLM), and CERES (CER)), for a) LH, b) SH, c) SW_d, d) SW_u, e) LW_d, f) LW_u, and g) R_{net}. For NEW, Tre, and Wil, the land mean has been approximated from published continental means as described in Section 2.b. Error bars are included where provided, for NEW and Wil these span the possible range described by multiple products, and for Jim and Mu1 these represent one standard deviation across multiple products (see citations for full details).

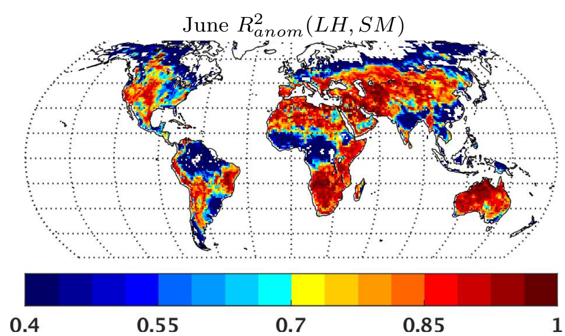

FIG. 2. R^2_{anom} between monthly anomalies of LH and rootzone soil moisture (SM) in MERRA-2 for JJA. No value is plotted where the correlation is negative.

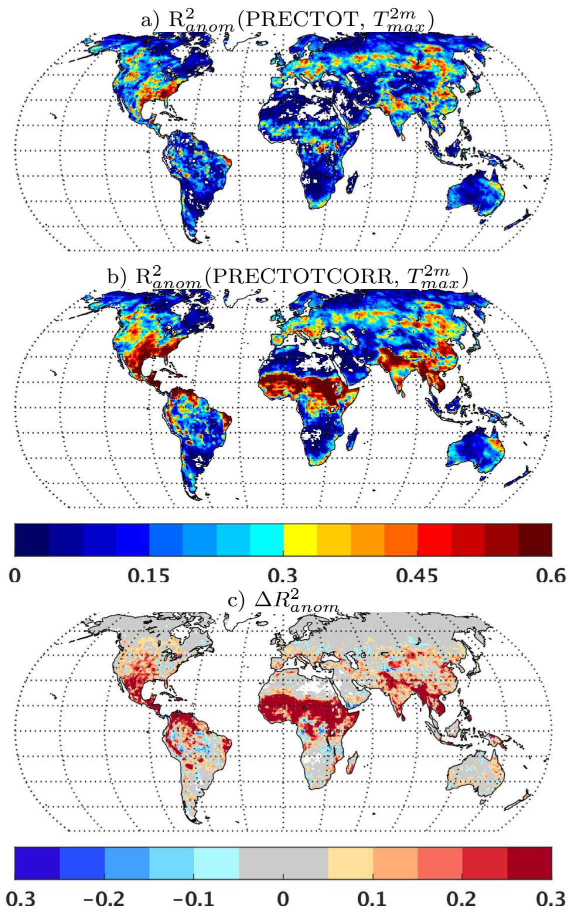

FIG. 3. JJA sensitivity of the monthly mean T^{2m}_{max} to precipitation in MERRA-2: R^2_{anom} between the monthly mean T^{2m}_{max} anomalies, and the two-monthly (current + previous months) precipitation anomalies, for (a) the model-generated precipitation (PRECTOT), and (b) the observation-corrected precipitation (PRECTOTCORR), together with their difference (c) $\Delta R^2_{anom} = R^2_{anom}(\text{PRECTOTCORR}, T^{2m}_{max})$ - $R^2_{anom}(\text{PRECTOT}, T^{2m}_{max})$. Values are plotted only where the correlation between T^{2m}_{max} and precipitation is negative.

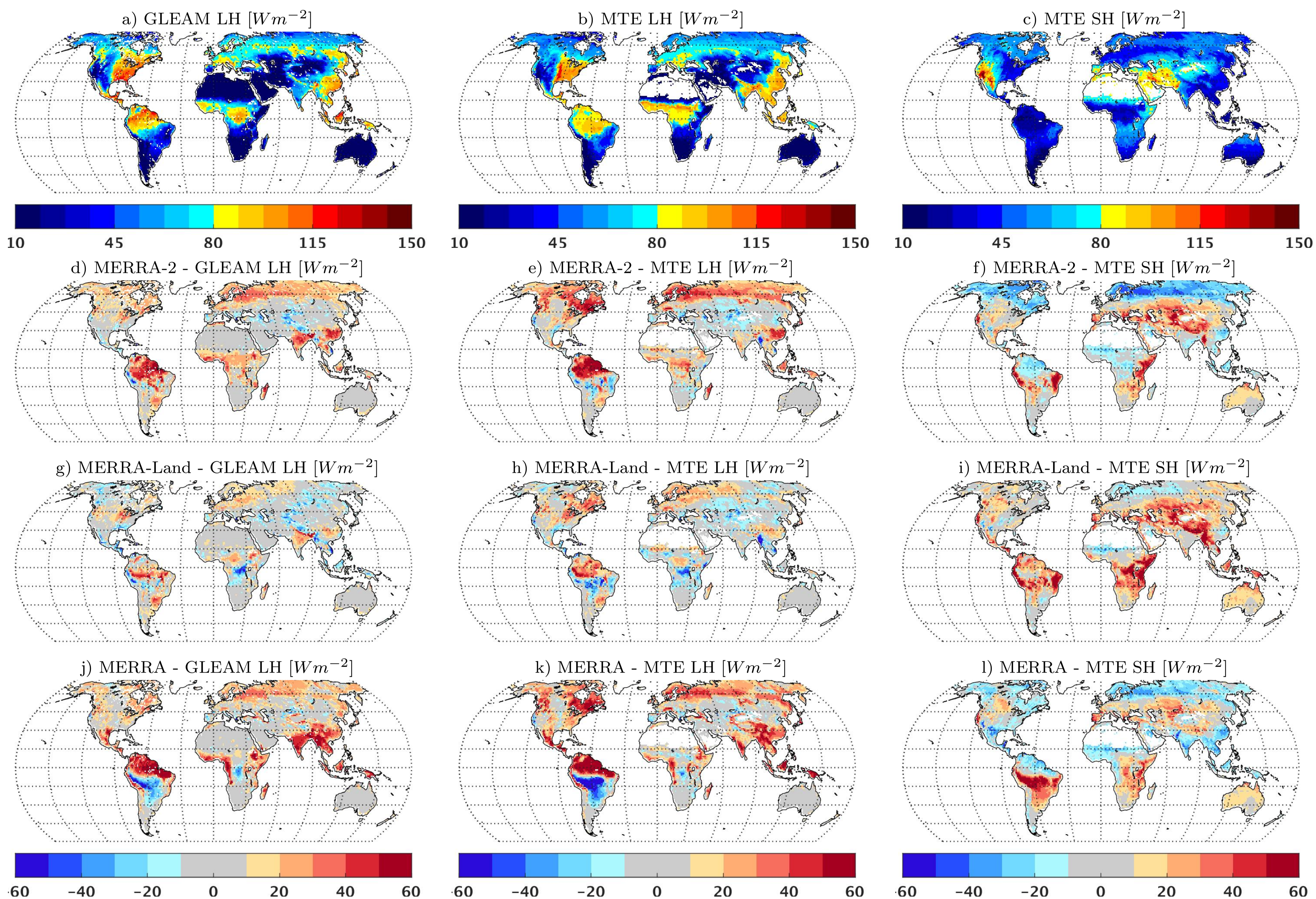

FIG. 4. The mean JJA turbulent fluxes, with GLEAM LH (column 1), MTE LH (column 2), and MTE SH (column 3) reference data in row 1, and the difference from the reference data for MERRA-2, MERRA-Land, and MERRA, in rows 2-4. The statistics span 1980-2016 for GLEAM, and 1982-2011 for MTE.

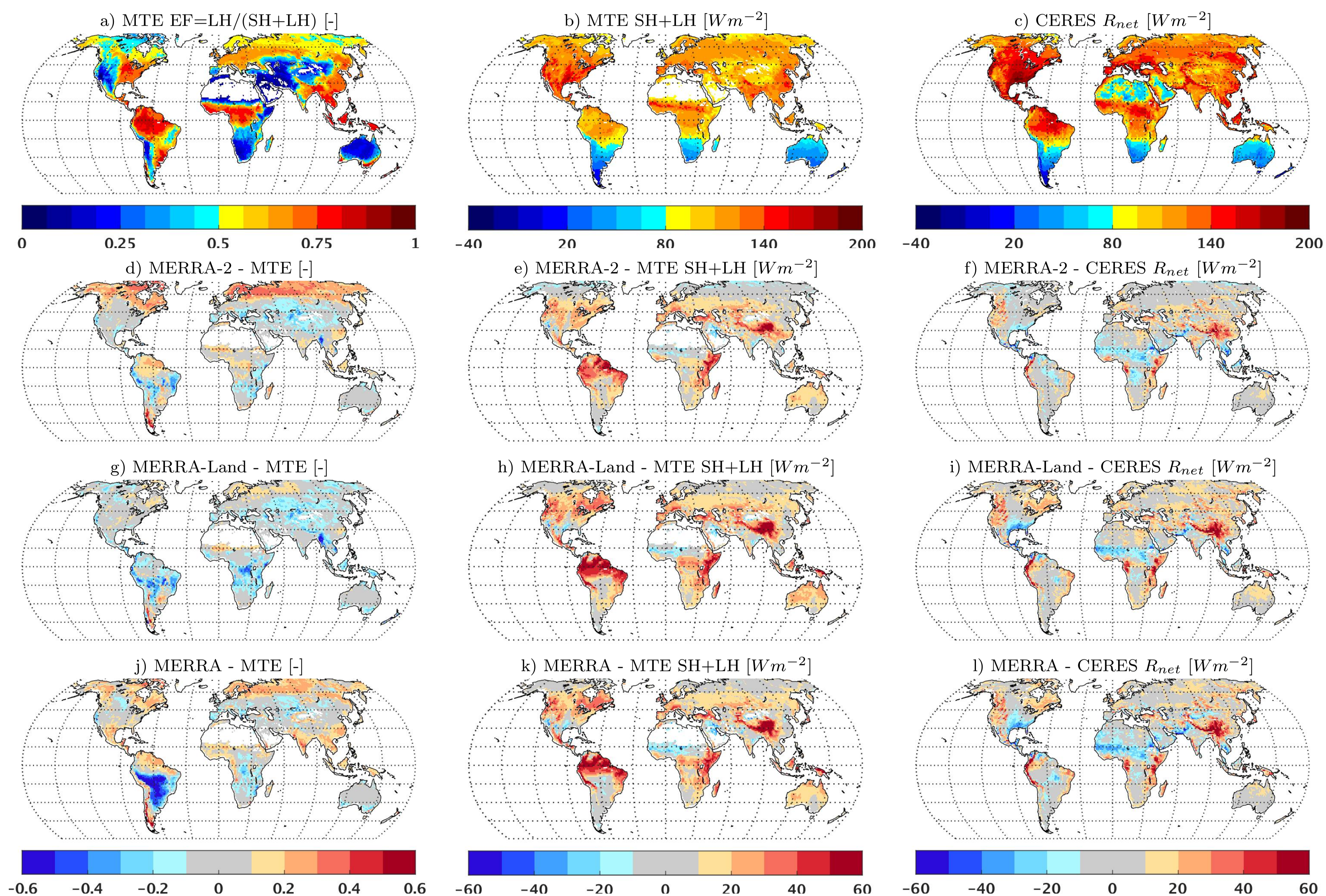

FIG. 5. Separation of mean JJA turbulent flux into evaporative fraction (EF) and incoming radiation biases, with the MTE EF (column 1), MTE LH+SH (column 2), and CERES-EBAF (column 3) reference data in row 1, and the difference from the reference data for MERRA-2, MERRA-Land, and MERRA in rows 2-4. The statistics span 1982-2011 for MTE, and 2000-2015 for CERES-EBAF.

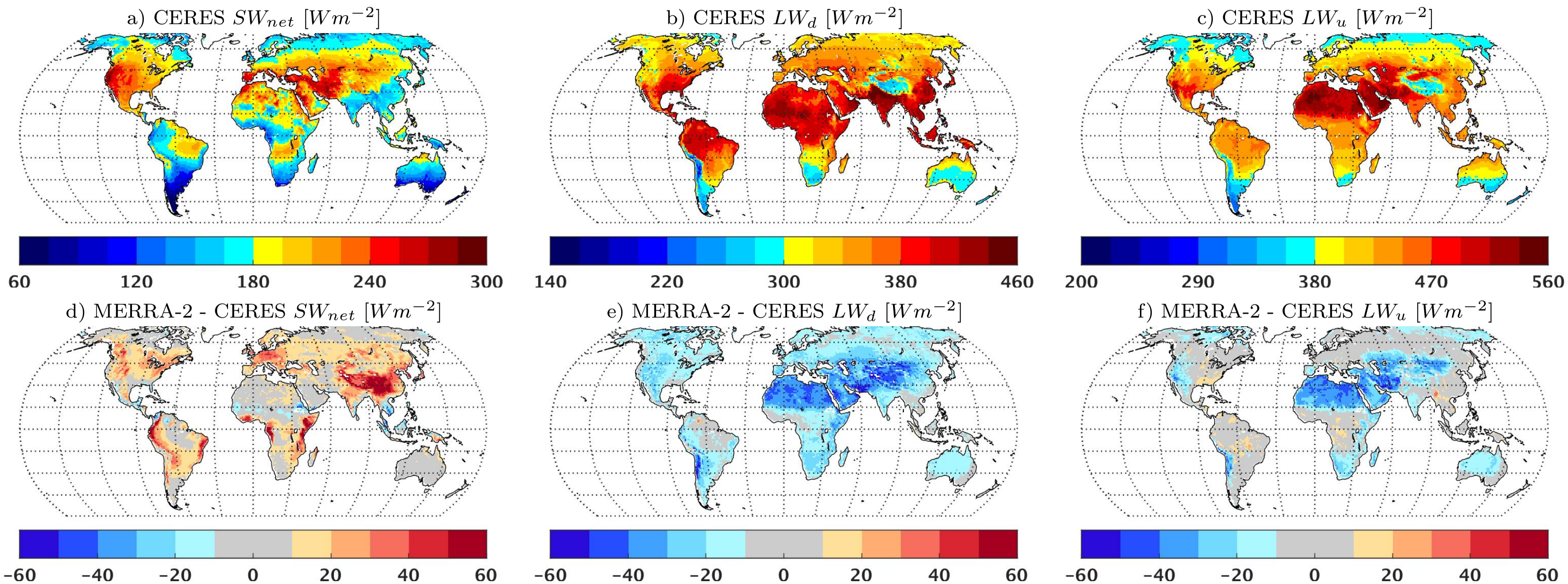

FIG. 6. The mean JJA radiation terms, from the CERES-EBAF reference data (row 1), and difference from the reference data for MERRA-2 (row 1), for (columns 1-3) SW_{net}, LW_u, and LW_d. The statistics span 2000-2015.

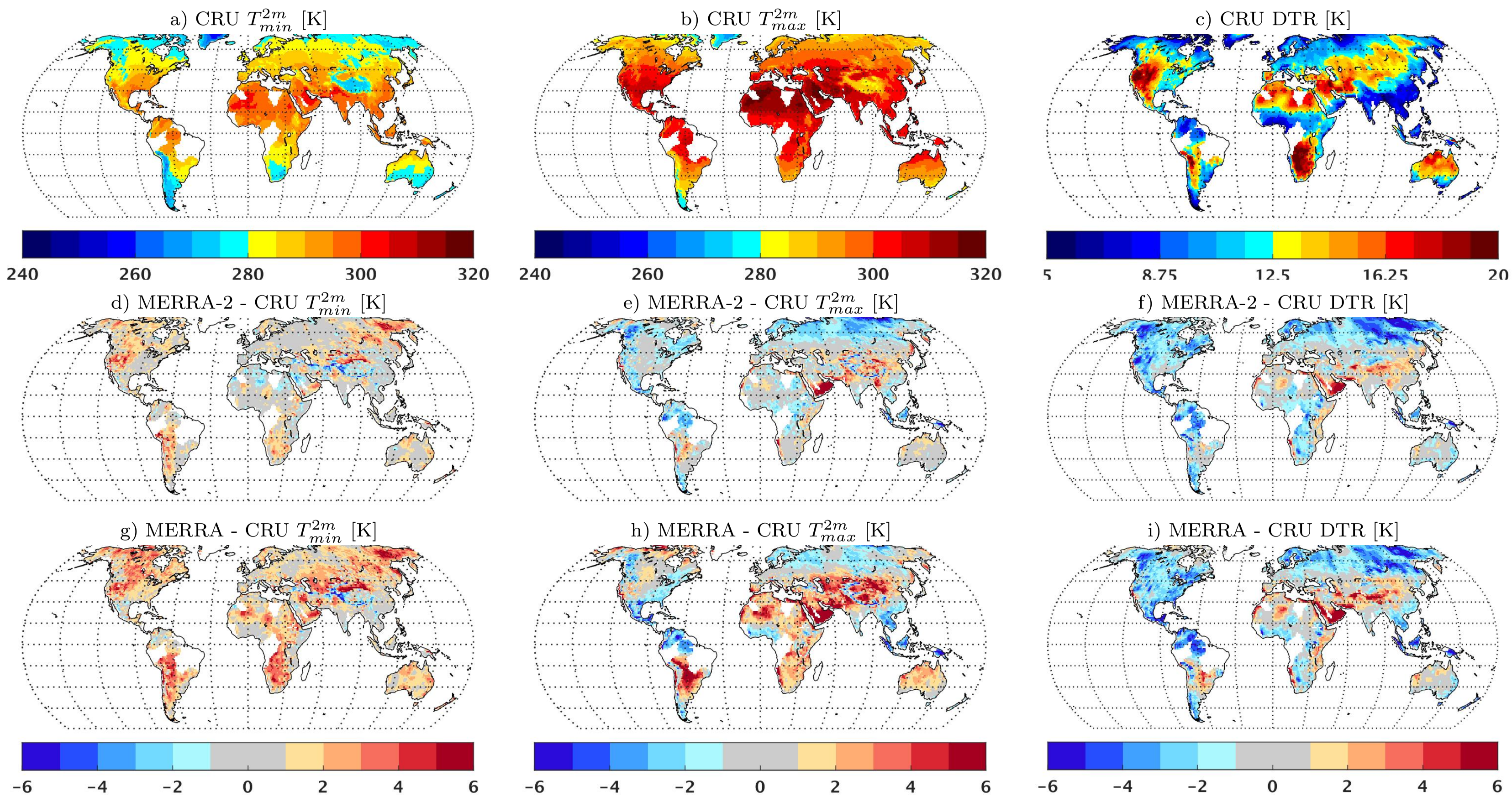

FIG. 7. The mean JJA T^{2m}, from CRU reference data (row 1), and the difference from the reference data for MERRA and MERRA-2 (rows 2-3), for the T^{2m}_{min} (column 1), T^{2m}_{max} (column 2), and the DTR (column 3). The statistics span 1980-2015, and white plotted over land indicates insufficient CRU data.

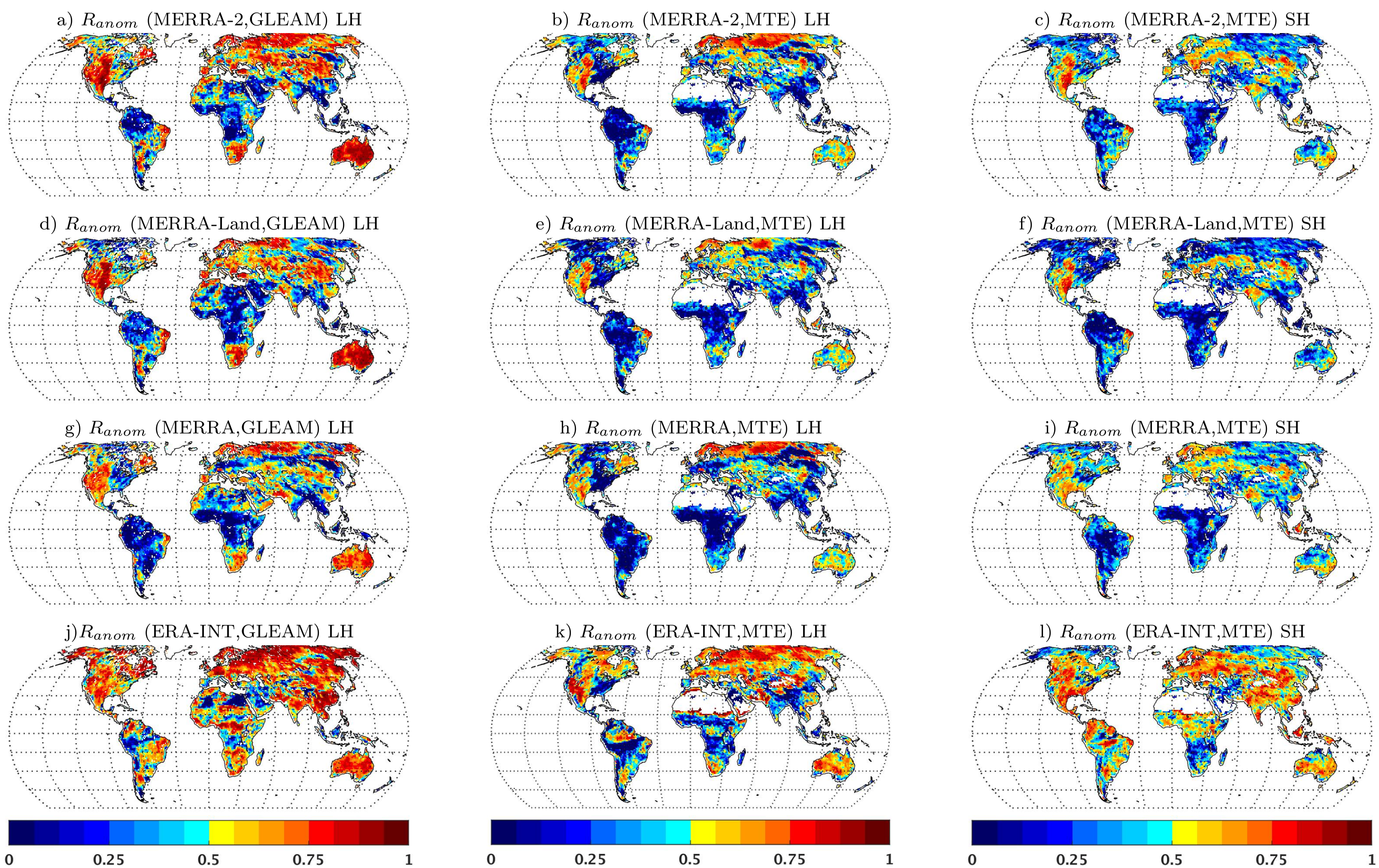

FIG. 8. R_{anom} between GLEAM LH, MTE LH, and MTE SH (columns 1-3) for MERRA, MERRA-Land, MERRA-2, and ERA-Interim (rows 1-4), for JJA. Statistics span 1980-2016 for GLEAM and 1982-2011 for MTE.

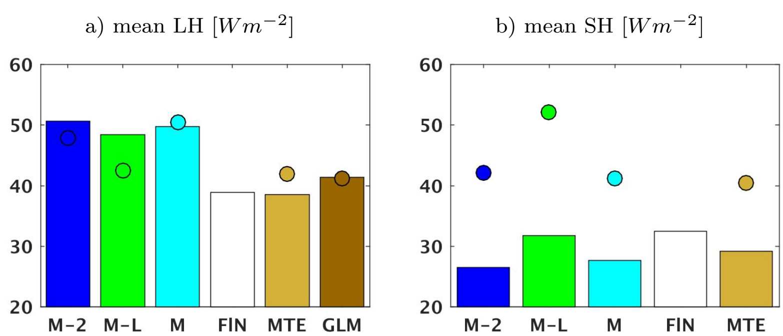

FIG. 9. Bar plot of the mean annual (a) LH and (b) SH, across the 20 Fluxnet site locations, from MERRA-2 (M-2), MERRA-Land (M-L), MERRA (M), Fluxnet (FlN), MTE, and GLEAM (GLM; LH only), calculated using each data set at its native resolution (and screened temporally for Fluxnet availability). For the global data sets, circles are plotted for the global land annual mean (taken from Figure 1).

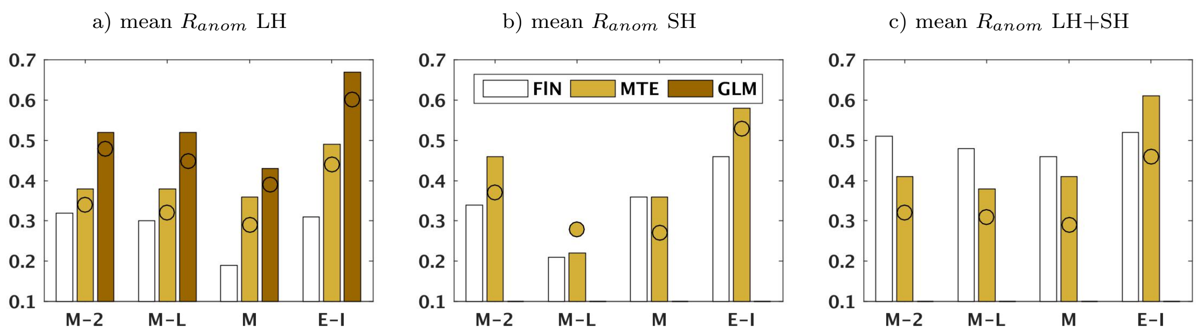

FIG. 10. Bar plot of the R_{anom} over JJA averaged across the 20 Fluxnet site locations, for (a) LH, (b) SH, and (c) LH+SH, between each pair of the reanalyses (MERRA-2 (M-2), MERRA-Land (M-L), MERRA (M), and ERA-I (E-I)) and the reference data (Fluxnet (FlN), MTE, and GLEAM (GLM)). The R_{anom} vs. the Fluxnet reference data use the reanalysis output at their reported spatial resolution (and screened temporally for Fluxnet availability), while the R_{anom} vs. GLEAM and MTE use reanalyses and reference data regridded to 1°. For GLEAM and MTE, circles are plotted for the global mean JJA R_{anom} (averaged over subplots of Figure 8).

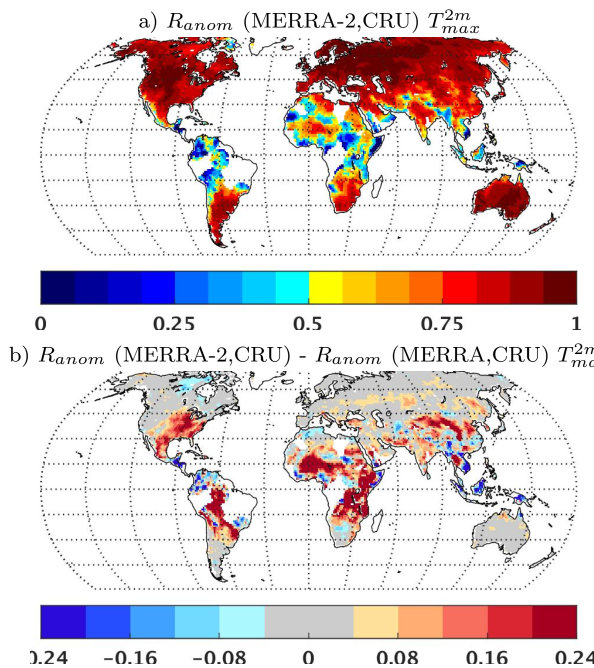

FIG. 11. The (a) MERRA-2 R_{anom} vs. CRU monthly mean T^{2m}_{max}, and (b) the improvement in the T^{2m}_{max} R_{anom} from MERRA to MERRA-2, both over JJA. Statistics span 1980-2015, and white plotted over land indicates insufficient CRU data.